Machine Intelligence Research
www.mi-research.net

22(5), October 2025, 817-848
DOI: 10.1007/s11633-025-1563-3

A Survey on Personalized Content Synthesis with Diffusion Models

Xulu Zhang[1,2] Xiaoyong Wei[1] Wentao Hu[1] Jinlin Wu[2,3] Jiaxin Wu[1]
Wengyu Zhang[1] Zhaoxiang Zhang[2,3,4] Zhen Lei[2,3,4] Qing Li[1]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong 999077, China

[2]Center for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences, Hong Kong 999077, China

[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

[4]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Recent advancements in diffusion models have significantly impacted content creation, leading to the emergence of personalized content synthesis (PCS). By utilizing a small set of user-provided examples featuring the same subject, PCS aims to tailor this subject to specific user-defined prompts. Over the past two years, more than 150 methods have been introduced in this area. However, existing surveys primarily focus on text-to-image generation, with few providing up-to-date summaries on PCS. This paper provides a comprehensive survey of PCS, introducing the general frameworks of PCS research, which can be categorized into test-time fine-tuning (TTF) and pre-trained adaptation (PTA) approaches. We analyze the strengths, limitations and key techniques of these methodologies. Additionally, we explore specialized tasks within the field, such as object, face and style personalization, while highlighting their unique challenges and innovations. Despite the promising progress, we also discuss ongoing challenges, including overfitting and the trade-off between subject fidelity and text alignment. Through this detailed overview and analysis, we propose future directions to further the development of PCS.

Keywords: Generative models, image synthesis, diffusion models, personalized content synthesis, subject customization.

Citation: X. Zhang, X. Wei, W. Hu, J. Wu, J. Wu, W. Zhang, Z. Zhang, Z. Lei, Q. Li. A survey on personalized content synthesis with diffusion models. *Machine Intelligence Research*, vol.22, no.5, pp.817–848, 2025. http://doi.org/10.1007/s11633-025-1563-3

1 Introduction

Recently, generative models have shown remarkable progress in the field of natural language processing and computer vision. Notable examples, such as ChatGPT[1] and text-to-image diffusion models[2], have showcased impressive capabilities in content creation. However, these advanced models often struggle to fulfill specific requirements, such as answering domain-specific queries or accurately depicting user′s portraits. This limitation highlights the importance of personalized content synthesis (PCS), which enables users to customize models for their unique tasks and requirements. As a critical area of research, PCS is increasingly recognized as essential for achieving artificial general intelligence (AGI), evidenced by the growing number of companies releasing products that support personalized content creation, like utilizing reinforcement learning to fine-tune language models[3]. In this paper, we focus on PCS within the context of diffusion models in computer vision. The objective of PCS is to learn the subject of interest (SoI) from a small set of user-uploaded samples and generate images that align with user-defined contexts. For instance, as illustrated in Fig. 1, PCS can customize a unique cat into various scenarios, such as "wearing pink sunglasses" and "in the snow".

The emergence of diffusion models has significantly facilitated text-guided content generation, leading to a rapid expansion of PCS. As depicted in Fig. 2, the number of research papers on PCS has surged over time. Following the release of key innovations such as DreamBooth[4] and textual inversion[7] in August 2022, over 150 methods have been proposed in a remarkably short period. These methods can be classified by using several criteria. First, in terms of training strategy, we differentiate between test-time fine-tuning (TTF) and pre-trained adaptation (PTA) approaches. TTF methods fine-tune generative

Review
Manuscript received on February 7, 2025; accepted on May 13, 2025
Recommended by Associate Editor Siwei Lyu
Colored figures are available in the online version at https://link.springer.com/journal/11633

Fig. 1 Given a few reference images of a subject (e.g., a cat[4] or face[5]), PCS aims to generate new renditions of the subject that align with user-defined textual prompts. The task requires preserving the subject′s identity while adapting to diverse contexts. The examples are generated by this survey using DreamBooth[4] and InstantID[6]. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

models for each personalization request during inference phase, while PTA methods strive to pre-train a unified model that is capable of generating a wide range of SoI. Second, we categorize the techniques employed in PCS into four primary areas: Attention-based operations, mask-guided generation, data augmentation and regularization. As shown in Fig. 2, we summarize various influential PCS methods based on these criteria. Third, the scope of personalization has expanded considerably. The SoI now not only includes general objects but also extends to human faces, artistic styles, actions and other intricate semantic elements. This growing interest encompasses the creation of complex compositions and the combination of several conditions. Additionally, the research landscape has broadened from static images to other modalities such as video and 3D representations.

While current methods in PCS have shown impressive performance, several challenges remain unresolved. A primary concern is the issue of overfitting, which often arises from the limited number of reference images available. This limitation can lead to a neglect of the textual context in the generated outputs, as shown in Fig. 3. Another related challenge is the trade-off between image alignment and text fidelity, as illustrated in Fig. 3. When a model successfully reconstructs the fine-grained details of the SoI, it often sacrifices controllability. Conversely, enhancing the model′s editability can result in underfitting with a loss of fidelity to the original SoI. Additionally, there are other challenges including the absence of robust evaluation metrics, a lack of standardized test datasets, and the need for faster processing times. This paper discusses these challenges in detail and establishes a benchmark for evaluating classical methods. By addressing these obstacles, we aim to advance the field of personalized content synthesis and enhance its practical applications.

This survey aims to provide a comprehensive overview of the existing methods, frameworks and challenges

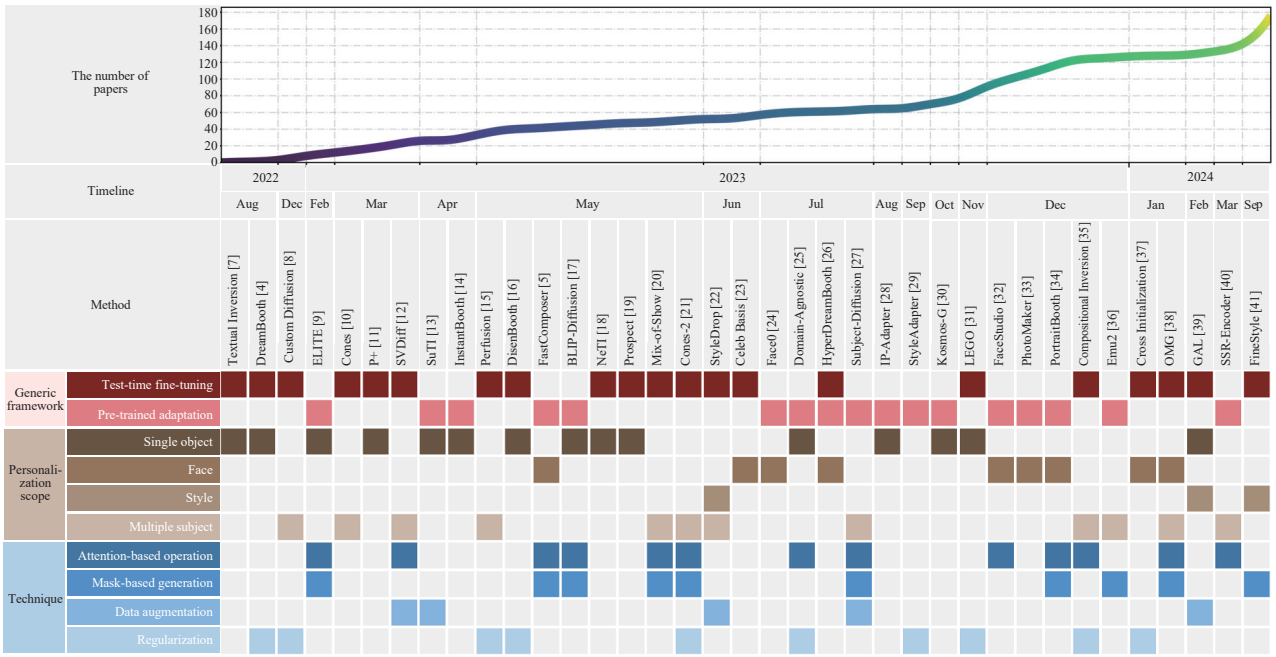

Fig. 2 A chronological overview of classical PCS methods as surveyed, illustrating the evolution of techniques through months. The number of related works has rapidly increased over the past two years. We divide PCS methods with 3 different criteria: training strategy, personalization scope, and technique. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

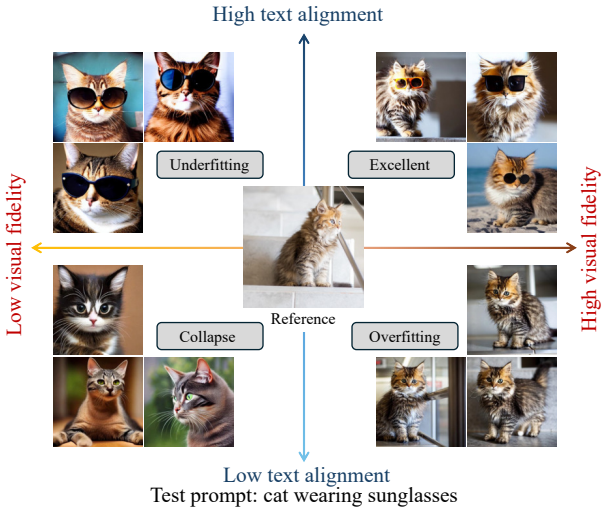

Fig. 3 The trade-off between text alignment and visual fidelity in personalized image synthesis, illustrated through DreamBooth-generated[4] examples of a customized cat wearing sunglasses. Overfitting occurs when the model focuses solely on reconstructing the cat, disregarding the sunglasses context. Underfitting, on the other hand, reflects the model's attempt to satisfy the text prompt but fails to accurately represent the personalized cat. Collapse signifies a failure to meet both criteria.

of PCS to help readers understand current trends in this evolving landscape and promote further improvements in this area. The structure of this survey is organized into several key sections: In Section 2, we begin with a brief introduction to diffusion models for better understanding of PCS. In Section 3, we introduce two primary frameworks for PCS: test-time fine-tuning and pre-trained adaptation. In Section 4, we categorize common techniques into four main parts, attention-based methods, mask-based generation, data augmentation and regularization. In Section 5, we summarize various PCS methods based on different image personalization tasks, including personalization on object, face, style, and etc. In Section 6, we demonstrate PCS's growing impact in video, 3D and other areas beyond traditional image generation tasks. In Section 7, we review the current evaluation dataset and metrics, and introduce a benchmark on the existing methods to promote further research. In Section 8, we discuss current challenges faced in PCS and propose potential future directions in this area. Additionally, we present a comprehensive summary of all related papers in Tables 1–3.

2 Fundamentals

Modern diffusion modeling has evolved into a sophisticated framework that unifies discrete and continuous generative paradigms through stochastic differential equations (SDEs)[180], and ordinary differential equations (ODEs)[180, 181]. This section introduces the core mathematical foundations, emphasizing innovations in the diffusion process and conditional mechanisms. By covering these theoretical developments, this section lays the groundwork for the personalization tasks.

2.1 Denoising diffusion probabilistic models (DDPMs)

The architecture of diffusion models comprises two complementary processes: A forward diffusion that systematically perturbs data distributions, and a learned reverse process that reconstructs signals through iterative refinement. Originating from discrete-time Markov chains in denoising diffusion probabilistic model (DDPM)[182], the forward process applies Gaussian noise corruption over T steps:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s \tag{1}$$

where $\mathbf{x}_t$ denotes the noised data at diffusion step t, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is the standard Gaussian noise vector that corrupts the original data $\mathbf{x}_0$, and α_s governs the noise scheduling policy.

The generative capability resides in learning to invert this degradation through parameterized reverse transitions. For DDPMs, this inversion is achieved via Bayesian reconstruction:

$$p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right), \beta_t I\right). \tag{2}$$

Here, p_θ represents the parameterized reverse transition probability, ϵ_θ is the neural network predicting noise components, and β_t is a hyperparameter before model training. And the neural network ϵ_θ learns to predict the applied noise component ϵ_θ through

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|_2^2\right]. \tag{3}$$

This objective enables end-to-end training of the denoising trajectory.

Once training is finished, the model can apply the learned reverse diffusion transitions to any arbitrary noise input, progressively denoising it into a coherent data sample. This capability allows DDPMs to function as general-purpose generative engines, producing high-quality outputs from random noise vectors.

2.2 SDEs

Although DDPMs establish basic denoising dynamics, this discrete formulation imposes constraints on flexible noise scheduling and computational efficiency.

To overcome these limitations, the framework was generalized through continuous-time SDE[180], which

Table 1 Paper summary on personalization

Paper	Scope	Framework	Backbone	Technique
Textual inversion[7]	Object	TTF	LDM	Token embedding fine-tuning
DreamBooth[4]	Object	TTF	Imagen	Diffusion model fine-tuning; regularization dataset
DreamArtist[42]	Object	TTF	LDM	Negative prompt fine-tuning
DVAR[43]	Object	TTF	SD 1.5	Randomness erasing
HiPer[44]	Object	TTF	SD	Token embedding enhancement
P+[11]	Object	TTF	SD 1.4	Token embedding enhancement
UNet-finetune[45]	Object	TTF	SD	Parameter-efficient fine-tuning
Jia et al.[46]	Object	TTF	Imagen	Mask-assisted generation; regularization
COTI[47]	Object	TTF	SD 2.0	Data augmentation
Gradient-free TI[48]	Object	TTF	SD	Evolutionary strategy
PerSAM[49]	Object	TTF	SD	Mask-assisted generation
DisenBooth[16]	Object	TTF	SD 2.1	Regularization
NeTI[18]	Object	TTF	SD 1.4	Token embedding enhancement
Prospect[19]	Object	TTF	SD 1.4	Token embedding enhancement
Break-a-scene[50]	Object	TTF	SD 2.1	Attention-based operation; attention-based operation; mask-assisted generation
COMCAT[51]	Object	TTF	SD 1.4	Attention-based operation
ViCo[52]	Object	TTF	SD	Attention-based operation; regularization
OFT[53]	Object	TTF	SD 1.5	Fine-tuning strategy
LyCORIS[54]	Object	TTF	SD 1.5	Attention-based operation
He et al.[55]	Object	TTF	SD, SDXL	Data augmentation
DIFFNAT[56]	Object	TTF	SD	Regularization
MATTE[57]	Object; style; high-level semantic	TTF	SD	Regularization
LEGO[31]	Object	TTF	SD	Regularization
Catversion[58]	Object	TTF	SD 1.5	Embedding concatenation
CLiC[59]	Object	TTF	SD 1.4	Attention-based operation; mask-assisted generation
HiFi tuner[60]	Object	TTF	SD 1.4	Attention-based operation; mask-assisted generation; regularization
InstructBooth[61]	Object	TTF	SD 1.5	Reinforcement learning
DETEX[62]	Object	TTF	SD 1.4	Mask-assisted generation
DreamDistribution[63]	Object	TTF	SD 2.1	Token embedding enhancement
DreamTuner[64]	Object	TTF	SD	Attention-based operation
Lu et al.[65]	Object	TTF	SD	Regularization; mask-assisted generation
Infusion[66]	Object	TTF	SD 1.5	Attention-based operation
Pair customization[67]	Object	TTF	SDXL	Disentanglement approach
DisenDreamer[68]	Object	TTF	SD 2.1	Disentanglement approach
DreamBooth++[69]	Object	TTF	SD	Attention-based operation
CoRe[70]	Object	TTF	SD 2.1	Regularization

Table 1 (continued) Paper summary on personalization

Paper	Scope	Framework	Backbone	Technique
DCI[71]	Object	TTF	SD	Token embedding optimization
Re-Imagen[72]	Object	PTA	Imagen	Retrieval-augmented paradigm
Versatile diffusion[73]	Object	PTA	SD 1.4	Architecture design
Tuning-encoder[74]	Object	PTA	SD	Regularization
ELITE[9]	Object	PTA	SD 1.4	Attention-based operation; mask-assisted generation
UMM-diffusion[75]	Object	PTA	SD 1.5	Reference feature injection
SuTI[13]	Object	PTA	Imagen	Data augmentation
InstantBooth[14]	Object	PTA	SD 1.4	Patch feature extraction
BLIP-diffusion[17]	Object	PTA	SD 1.5	Data augmentation; mask-assisted generation
Domain-agnostic[25]	Object	PTA	SD	Regularization; attention-based operation
IP-Adapter[28]	Object	PTA	SD 1.5	Reference feature injection
Kosmos-G[30]	Object; face	PTA	SD 1.5	Multimodal language modeling
Kim et al.[76]	Object	PTA	SD 1.4	Reference feature injection
CAFÉ[77]	Object; face	PTA	SD 2.0	Multimodal language modeling
SAG[78]	Object; face; style	PTA	SD 1.5	Gradient-based guidance
BootPIG[79]	Object	PTA	SD	Data augmentation
CustomContrast[80]	Object	PTA	SD 1.5	Contrastive learning
MoMA[81]	Object	PTA	SD 1.5	Multimodal LLM adapter
Diptych prompting[82]	Object; style	PTA	FLUX	Attention-based operation; mask-assisted generation
StyleDrop[22]	Style; multiple subjects	TTF	Muse	Data augmentation
StyleBoost[83]	Style	TTF	SD 1.5	Regularization dataset
StyleAligned[84]	Style	TTF	SDXL	Regularization
GAL[39]	Style; object	TTF	SD 1.5	Data augmentation
StyleForge[85]	Style	TTF	SD 1.5	Parameter-efficient fine-tuning
FineStyle [41]	Style	TTF	Muse	Mask-assisted generation
Style-friendly[86]	Style	TTF	FLUX-dev; SD 3.5	SNR sampler
StyleAdapter[29]	Style	PTA	SD	Data augmentation; reference feature extraction
ArtAdapter[87]	Style	PTA	SD 1.5	Data augmentation
ProFusion[88]	Face	TTF	SD 2.0	Fusion sampling
Celeb basis[23]	Face	TTF	SD	Face feature representation
Banerjee et al.[89]	Face	TTF	SD 1.4	Regularization
Magicapture[90]	Face; multiple subjects	TTF	SD 1.5	Attention-based operation; mask-assisted generation; identity loss
Concept-centric[91]	Face	TTF	SD 1.5	Modified classifier-free guidance
Cross initialization[37]	Face	TTF	SD 2.1	Regularization

Table 2 Paper summary on personalization

Paper	Scope	Framework	Technique	Backbone
OMG[38]	Face; multiple subjects	TTF	SDXL	Mask-assisted generation; regularization
InstantFamily[92]	Face; multiple subjects; extra conditions	TTF	SD 1.5	Mask-assisted generation
PersonaMagic[93]	Face	TTF	SD 1.4	Attention-based operation
HyperDreamBooth[26]	Face	TTF; PTA	SD 1.5	Pretraining and fast fine-tuning
Su et al.[94]	Face	PTA	LDM; StyleGAN	Attention-based operation; regularization
FastComposer[5]	Face	PTA	SD 1.5	Object; style; high-level semantic
Face0[24]	Face	PTA	SD 1.4	Fusion sampling
DreamIdentity[95]	Face	PTA	SD 2.1	Attention-based operation
Face-diffuser[96]	Face	PTA	SD 1.5	Attention-based operation; mask-assisted generation
W-plus-adapter[97]	Face	PTA	SD 1.5; StyleGAN	Face feature representation
Portrait diffusion[98]	Face	PTA	SD 1.5	Mask-assisted generation
RetriNet[99]	Face	PTA	SD	Mask-assisted generation
FaceStudio[32]	Face	PTA	SD	Attention-based operation
PVA[100]	Face	PTA	LDM	Mask-assisted generation
DemoCaricature[101]	Face; extra conditions	PTA	SD 1.5	Regularization
PhotoMaker[33]	Face	PTA	SDXL	Regularization; mask-assisted generation
Stellar[102]	Face	PTA	SDXL	Evaluation prompts and metrics
PortraitBooth[34]	Face	PTA	SD 1.5	Attention-based operation; mask-assisted generation
InstantID[6]	Face; extra conditions	PTA	SDXL	Multi-feature injection
ID-Aligner[103]	Face	PTA	SD 1.5; SDXL	Feedback learning
MoA[104]	Face; multiple subjects	PTA	SD 1.5	Attention-based operation; mask-assisted generation
IDAdapter[105]	Face	PTA	SD 2.1	Mixed facial features
Infinite-ID[106]	Face	PTA	SDXL	Reference feature injection
Face2Diffusion[107]	Face	PTA	SD	Multi-feature injection
FreeCure[108]	Face	PTA	SD 1.5; SDXL	Mask-assisted generation
Omni-ID[109]	Face	PTA	FLUX	Face representation enhancement
Custom diffusion[8]	Multiple subjects	TTF	SD 1.4	Style; multiple subjects
Cones[10]	Multiple subjects	TTF	SD 1.4	Concept neurons activation
SVDiff[12]	Multiple subjects	TTF	SD	Data augmentation; attention-based operation
Perfusion[15]	Multiple subjects	TTF	SD 1.5	Regularization
Mix-of-show[20]	Multiple subjects	TTF	SD 1.5	Data augmentation; reference feature extraction
Cones-2[21]	Multiple subjects	TTF	SD 2.1	Face; multiple subjects

Table 2 (continued) Paper summary on personalization

Paper	Scope	Framework	Technique	Backbone
PACGen[110]	Multiple subjects	TTF	SD 1.4	Mask-assisted generation
Compositional inversion[35]	Multiple subjects	TTF	SD	Attention-based operation; mask-assisted generation; identity loss
EM-style[111]	Multiple subjects	TTF	SD	Face; multiple subjects
MC^2[112]	Multiple subjects	TTF	SD 1.5	Attention-based operation
MultiBooth[113]	Multiple subjects	TTF	SD 1.5	Mask-assisted generation
Matsuda et al. [114]	Multiple subjects	TTF	LDM	Mask-assisted generation
MagicTailor[115]	Multiple subjects	TTF	SD 2.1	Attention-based operation; mask-assisted generation; identity loss
AnyDoor[116]	Multiple subjects	PTA	SD 2.1	Mask-assisted generation
Subject-diffusion[27]	Multiple subjects	PTA	SD 2.0	Face; multiple subjects; extra conditions
CustomNet[117]	Identity loss; multi-task learning	PTA	SD	Multi-condition integration
MIGC[118]	Multiple subjects	PTA	SD	Attention-based operation; mask-assisted generation
Emu2[36]	Multiple subjects	PTA	SDXL	In-context learning
SSR-encoder[40]	Multiple subjects	PTA	SD 1.5	Attention-based operation
λ-eclipse[119]	Multiple subjects	PTA	Kandinsky v2.2	Contrastive learning
MS-diffusion[120]	Multiple subjects	PTA	SD	Attention-based operation
ReVersion[121]	High-level semantic	TTF	SD	Regularization
Inv-ReVersion[122]	High-level semantic	TTF	SD 1.5	Regularization
CusConcept[123]	High-level semantic	TTF	SD 2.1	Regularization
ADI[124]	High-level semantic	TTF	SD 2.1	Mask-assisted generation
PhotoSwap[125]	Extra conditions	TTF	SD 2.1	Attention-based operation
PE-VITON[126]	Extra conditions	TTF	SD	Shape and texture control
Layout-control[127]	Extra conditions	TTF	SD	Attention-based operation
SwapAnything[128]	Extra conditions	TTF	SD 2.1	Mask-assisted generation
PE-VITON[126]	Extra conditions	TTF	SD	Shape and texture control
Viewpoint control[129]	Extra conditions	TTF	SDXL	3D feature incorporation
Prompt-free diffusion[130]	Extra conditions	PTA	SD 2.0	Reference feature injection
Uni-ControlNet[131]	Extra conditions	PTA	SD	Multi-feature injection
ViscoNet[132]	Extra conditions	PTA	SD	Multi-feature injection
Context diffusion[133]	Extra conditions	PTA	LDM	Multi-feature injection
FreeControl[134]	Extra conditions	PTA	SD 1.5; SD 2.1; SDXL	Fusion guidance
Li et al.[135]	Extra conditions	PTA	SD	Face; extra conditions

provides a unified perspective for modeling diffusion dynamics:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \tag{4}$$

where $\mathbf{f}(\mathbf{x}, t)$ encodes deterministic drift components, $g(t)$ modulates stochastic diffusion via Wiener process $\mathbf{w}$. This continuous perspective subsumes DDPMs as special cases while enabling adaptive noise scheduling strategies such as variance-preserving (VP) and variance-exploding (VE) schedules through strategic choices of $\mathbf{f}$ and g[181].

Table 3 Paper summary on personalization

Paper	Scope	Framework	Technique	Backbone
Tune-A-video[136]	Video	TTF	SD 1.4	Parameter-efficient fine-tuning
Gen-1[137]	Video	TTF	LDM	Multi-feature injection
Make-A-protagonist[138]	Video	TTF	LDM	Attention-based operation
Animate-A-story[139]	Video	TTF	SD	Retrieval-augmented paradigm
MotionDirector[140]	Video	TTF	Zeroscope; modelscope	Disentanglement approach
LAMP[141]	Video	TTF	SDXL; SD 1.4	First-frame conditioned pipeline
VMC[142]	Video	TTF	Show-1	Parameter-efficient fine-tuning
SAVE[143]	Video	TTF	VDM	Attention-based operation
Customizing motion[144]	Video	TTF	ZeroScope	Regularization
DreamVideo[145]	Video	TTF	ModelScope	Disentanglement approach
MotionCrafter[146]	Video	TTF	VDM	Disentanglement approach
Customize-A-video[147]	Video	TTF	ModelScope	Multi-stage refinement
Magic-Me[148]	Video	TTF	SD 1.5; AnimateDiff	Multi-stage refinement
CustomTTT[149]	Video	TTF	VDM	Fine-tuning and alignment
PersonalVideo[150]	Video; face	TTF	AnimateDiff	Identity injection
CustomVideo[151]	Video; multiple subjects	TTF	ZeroScope	Attention-based operation; regularization
VideoDreamer[152]	Video; multiple subjects	TTF	SD 2.1	Data augmentation
VideoAssembler[153]	Video	PTA	VidRD	Reference feature injection
VideoBooth[154]	Video	PTA	VDM	Reference feature injection
VideoMaker[155]	Video	PTA	VDM	Reference feature injection
SUGAR[156]	Video	PTA	VDM	Attention-based operation
StyleCrafter[157]	Video; style	PTA	SD 1.5	Reference feature injection
ID-Animator[158]	Video; face	PTA	AnimateDiff	Reference feature injection
ConsisID[159]	Video; face	PTA	CogVideoX	Frequency decomposition
MotionCharacter[160]	Video; face	PTA	VDM	Attention-based operation; mask-assisted generation
DreaMoving[161]	Video; face; extra conditions	PTA	SD	Multi-feature injection
Magic3D[162]	3D	TTF	Imagen; LDM	Score distillation sampling
DreamBooth3D[163]	3D	TTF	Imagen	Multi-stage refinement
PAS[164]	3D	PTA	SD	Text-to-3D-pose
StyleAvatar3D[165]	3D	PTA	LDM	Multi-view alignment
AvatarBooth[166]	3D	TTF	SD	Dual model fine-tuning
MVDream[167]	3D	TTF	SD 2.1	Score distillation sampling
Consist3D[168]	3D	TTF	SD	Token embedding enhancement
Animate124[169]	3D	TTF	SD 1.5	Multi-stage refinement

Table 3 (continued) Paper summary on personalization

Paper	Scope	Framework	Technique	Backbone
Dream-in-4D[170]	3D	TTF	SD 2.1	Multi-stage refinement
TextureDreamer[171]	3D	TTF	LDM	Texture extraction
TIP-editor[172]	3D	TTF	SD	Regularization; mask-assisted generation
Anti-DreamBooth[173]	Attack and defense	TTF	SD 2.1	Perturbation learning
Concept censorship[174]	Attack and defense	TTF	SD 1.4	Trigger injection
Huang et al.[175]	Attack and defense	TTF	SD	Backdoor attack
Continual diffusion[176]	Others	TTF	SD	Continual learning
SVGCustomization[177]	Others	TTF	SD 1.5	Fine-tuning and alignment
StitchDiffusion[178]	Others	TTF	LDM	Parameter-efficient fine-tuning
MC-TI[179]	Others	TTF	SD 1.5	Regularization

For the reverse process, a crucial result[183] shows that the reverse diffusion process follows an SDE:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \quad (5)$$

where $\bar{w}$ denotes reverse-time Wiener increments, and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is a score function. This formulation enables generation by integrating backward from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ to $\mathbf{x}_0$, conditioned on learned score estimates.

2.3 ODEs

While SDEs provide a comprehensive framework for modeling diffusion dynamics, their inherent randomness introduces critical challenges in practical deployment. The Wiener process $\mathbf{w}$ in SDE-based sampling generates pathwise variability, requiring extensive Monte Carlo averaging to achieve stable solutions. Furthermore, stochastic trajectories exhibit erratic curvature profiles, forcing fixed-step solvers to adopt conservative discretization schemes that often necessitate 1 000+ steps for high-fidelity generation.

These limitations motivate the derivation of deterministic sampling trajectories through probability flow ODEs[180, 181], obtained by eliminating the stochastic term from the SDE formulation:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t. \quad (6)$$

The ODE's deterministic nature arises from two synergistic components: The original drift term $\mathbf{f}(\mathbf{x}, t)$ and a correction term involving the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$. This score-guided adjustment preserves the marginal data distribution $p_t(\mathbf{x})$ while eliminating pathwise stochasticity. Crucially, the resulting trajectories exhibit intrinsic geometric regularity, allowing the complex denoising process to operate in a simplified mathematical space. This enables 5–10 times faster sampling than SDE-based methods by leveraging adaptive ODE solvers such as DPM-Solver[184] and DPM-Solver++[185]. Recent work[186] further identifies strong geometric regularity in these ODE-based sampling trajectories, demonstrating that they inherently follow a linear-nonlinear-linear structure regardless of generated content. This trajectory regularity enables dynamic programming-based time scheduling optimization with negligible computational overhead.

2.4 Conditional generation mechanisms

Recently, conditional synthesis has emerged as the critical capability bridging theoretical diffusion frameworks with real-world applications. It enables precise alignment of outputs with multimodal guidance signals (text prompts, subject embeddings, anatomical masks), fulfilling domain-specific requirements for reliability and reproducibility.

Building upon the unconditional framework, conditional synthesis is formalized through extended score matching:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, t, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, c)\|_2^2\right] \quad (7)$$

where the conditioning signal c can be integrated via multiple synergistic mechanisms, such as cross-modal attention[187] and spatial modulation[188].

This conditional paradigm directly enables the applications of PCS. Leading text-to-image systems, such as stable diffusion (SD)[187] and DALLE[189], are widely adopted to empower users in controlling customized content through text instructions.

3 Generic framework

In this survey, we broadly categorize PCS frameworks into two paradigms: TTF and PTA approaches. These two frameworks fundamentally differ in their adaptation mechanisms. TTF methods dynamically adjust

model parameters for each new subject during inference, prioritizing visual fidelity at the cost of computational overhead. Conversely, PTA frameworks employ reference-aware architectures trained on large datasets to enable single-pass personalization without parameter updates during inference. We introduce two generic frameworks in Sections 3.1 and 3.2.

3.1 TTF framework

TTF methods represent the foundational approach for personalized synthesis, where models adapt to new subjects through instance-specific optimization during inference. As illustrated in Fig. 4, this framework operates through two core principles: 1) **Test-time adaptation** fine-tunes model parameters to learn the key visual element of the SoI. 2) **Semantic-aware modifier system** represents the SoI at the token level to bridge the gap between visual adaptation and textual control. In Sections 3.1.1 and 3.1.2, we detail these principles respectively.

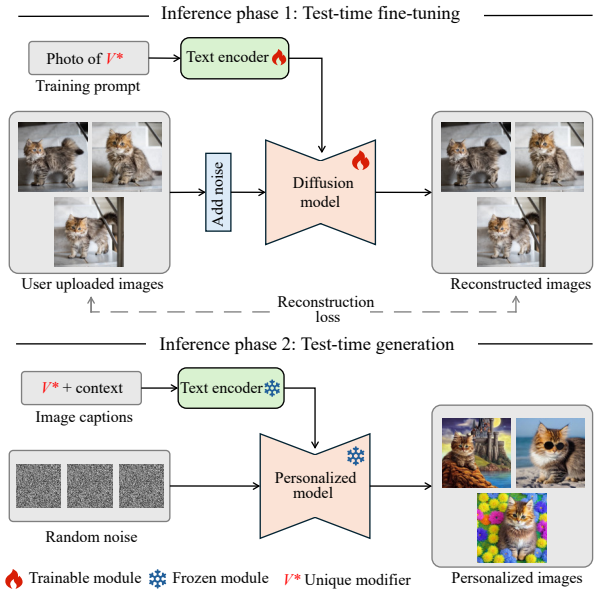

Fig. 4 Illustration of the TTF framework for the test-time fine-tuning process and generation phase. During the inference phase, the model fine-tunes its parameters by reconstructing the reference images for each SoI group. The unique modifier V^* is employed to represent the SoI and used to formulate new inference prompts for generating personalized images. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

3.1.1 Test-time adaptation

Test-time fine-tuning. For each reference set $\mathbf{X}_{SoI}$ of SoI, the optimization process adjusts a subset of parameters $\boldsymbol{\theta}'$ to reconstruct the SoI conditioned on the reference prompt. The fine-tuning objective is defined by a reconstruction loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0 \in \mathbf{X}_{SoI}, t, \boldsymbol{\epsilon}} \left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta'}(\mathbf{x}_t, t, c) \|_2^2 \right] \tag{8}$$

where c represents the condition signal, typically the reference image caption. Compared to the large-scale pre-training described in (7), key differences lie in the training data and learnable parameters. The training samples are typically restricted to the SoI references, sometimes supplemented with a regularization dataset to mitigate overfitting[4]. For the selection of learnable parameters $\boldsymbol{\theta}'$, the commonly adopted options include token embeddings[7, 11], the entire diffusion model[4, 39], and specific subsets of parameters[8, 12, 15], or include introducing new parameters such as adapters[22, 45] and low-rank adaptation (LoRA)[20, 26, 38], which will be discussed in Section 3.1.3.

Test-time generation. The test-time generation is proceeded once the model has been fine-tuned with the optimized parameters $\boldsymbol{\theta}'$. By composing novel input prompts that incorporate the SoI's unique identifier (introduced in Section 3.1.2), the adapted model can synthesize diverse images while preserving the subject's distinctive characteristics. This approach enables flexible control over the generated content through natural language descriptions.

3.1.2 Unique modifier

A unique modifier is a textual token or short phrase that uniquely represents an SoI, allowing it to be referenced textually in prompts. As illustrated in Fig. 4, this modifier serves as a textual description for the SoI, allowing combination with other descriptions (e.g., "V^* on the beach") during inference. Normally, the construction of the unique modifier can be divided into three categories.

Plain text[4, 8, 15, 22, 39]. This approach utilizes an explicit text description to represent the SoI. For example, words such as cat could directly represent the user's cat in the references. This setting typically requires fine-tuning the parameters of the diffusion model components, such as UNet[190] or transformer blocks[191], to enable the model to associate the SoI's visual features with the plain-text token. The plain text offers user-friendly prompt construction and injects subject prior information to ease tuning difficulty. However, this technique may over-specialize common terms, limiting their broader applicability as the model learns to associate general words with specific SoI characteristics.

Rare token[4]. This technique employs infrequently used tokens to minimize their impact on commonly used vocabulary. Similar to the plain text approach, the embeddings of rare tokens remain unchanged during fine-tuning. However, these rare tokens often fail to provide useful subject prior information and still exhibit weak interference with unrelated vocabulary, potentially leading to ambiguity between the original meaning and the intended SoI reference.

Learnable token embedding[7, 11, 18, 19, 37, 60]. This method adds a new token and its corresponding embedding vector to the dictionary of the tokenizer. An intuitive example is that this method creates a new word that

does not exist in the dictionary. This inserted token has adjustable weights during fine-tuning, while the embeddings of other tokens in the pre-defined dictionary will remain unchanged. This approach requires only a few kilobytes of additional parameters and maintains the base model's capabilities for non-customized generation. Similar to the rare token approach, it is not very user-friendly in practice, since users must learn and remember an unfamiliar token to reference their subject of interest.

3.1.3 Training parameter selection

The selection of trainable parameters represents a critical design consideration in PCS, directly impacting many key performance metrics: Subject fidelity, training efficiency and model storage requirements. Current tunable parameters can be broadly categorized into the following four types.

Token embedding. As introduced in Section 3.1.2, token embedding optimization[7] introduces learnable tokens as the unique modifier to represent SoI through noise-to-image reconstruction. While achieving remarkable parameter efficiency, this approach faces challenges in detail preservation and prolonged training time (typically longer than 20 minutes) due to the inherent compression of complex features into low-dimensional embeddings. Subsequent works[11, 18, 19, 37, 60] aim to address these limitations through different strategies, which are summarized in Section 5.1.

Existing model parameters. This paradigm directly optimizes pre-trained model components, such as the text encoder, UNet blocks and transformer layers[4, 8, 12, 15, 39]. Benefitting from the advanced representation capacity of these modules, the fine-tuning phase can achieve faster convergence (5–10 minutes) and superior visual fidelity compared to token-only methods, though at the cost of significant storage overhead. In addition, these modules inherently support attention mechanisms to facilitate feature enhancement operations.

Parameter-efficient extensions. Recent advanced methods have introduced parameter-efficient techniques into PCS, such as LoRA[20, 26, 38, 54, 65, 102] and adapter modules[22, 45], which inject small, trainable components into the base model. These methods achieve comparable performance to full parameter fine-tuning while dramatically reducing storage requirements.

Combined strategy. Since the aforementioned strategies are not conflicting, some methods allocate different learning rates and training phases to each component type to achieve optimal balance between fidelity and efficiency. For instance, the fine-tuned token embedding can be regarded as an effective initialization for the subsequent model weight fine-tuning[90]. Also, these two parts can be simultaneously optimized with different learning rates[50, 64].

3.1.4 Prompt engineering

The construction of training prompts for samples typically starts with adding prefix words before the modifier token. A simplest example is "photo of V^*". However, DreamBooth[4] notes that such a simple description causes a long training time and unsatisfactory performance. To address this, they incorporate the unique modifier with a class noun to describe the SoI in the references (e.g., "photo of V^* cat"). Also, the training caption for each training reference can be more precious for better disentanglement of the SoI and irrelevant concepts[55], such as "photo of V^* cat on the chair". This follows the trend that high-quality captions in the training set could assist in further improvement of accurate text control[192].

3.2 PTA framework

The PTA framework has emerged as a breakthrough approach for PCS, aiming to eliminate the computational burden of per-request fine-tuning while maintaining high-quality and subject-specific generation capabilities. To achieve this goal, this approach combines large-scale pretraining with reference-aware architectures to enable single-pass personalization, as shown in Fig. 5. Based on this architecture, three critical design factors are considered to ensure practical viability: 1) **Preservation of semantic-critical features** to guarantee visual consistency with reference inputs, 2) **effective fusion** that combines reference features with text guidance to achieve desired generation, 3) **optimization of training dataset** scale to achieve robust generalization without overfitting. In this section, we first present overall architecture introduction in Section 3.2.1 and then delve into three detailed factors in Sections 3.2.2–3.2.4.

3.2.1 Pre-trained adaption

Pre-training. During the pre-training phase, the PTA framework aims to establish direct mappings between reference characteristics (e.g., facial features, object textures) and synthesized outputs. To achieve this, the reference inputs are processed through dedicated feature extractors and are fused with text prompts to serve as conditions to guide the generation, as shown in Fig. 5. A reconstruction loss is optimized that enforces the alignment between generated images and the large-scale dataset X_{data}:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0 \in \mathbf{X}_{data}, t, \boldsymbol{\epsilon}} \left[\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta'}(\mathbf{x}_t, t, c') \|_2^2 \right], \quad c' = \mathcal{F}(c, \mathbf{x}_0) \tag{9}$$

where $\mathcal{F}$ represents the fusion operation combining text condition c and reference image $\mathbf{x}_0$. The tunable parameters $\boldsymbol{\theta}'$ include visual encoder weights, text encoder components, diffusion modules and injected adapter modules.

Inference. During inference, the PTA framework processes a reference image through its visual encoder to extract discriminative features, which are then fused with text embeddings using the pretrained conditioning module. This fused representation guides the diffusion model to generate personalized outputs. This approach effect-

ively eliminates test-time optimization to ensure fast generation.

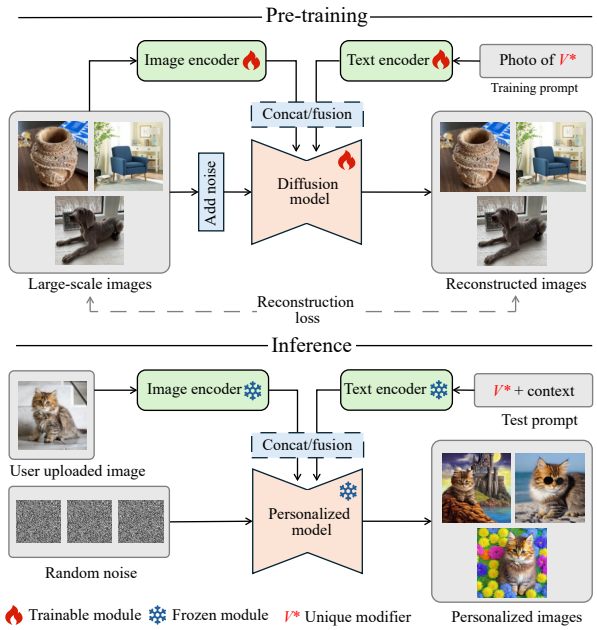

Fig. 5 Illustration of the PTA method for personalized image synthesis. This framework utilizes a large-scale dataset to train a unified model that can process diverse personalization requests. The diffusion model is adapted to process hybrid inputs derived from both visual and textual features. (Colored figures are available in the online version at https://link.springer.com/journal/11633)

3.2.2 Subject feature extraction

Extracting representative features of the SoI is crucial in the creation of personalized content. A common approach is to employ an encoder, leveraging pre-trained models such as CLIP[193] and BLIP[194]. While these models excel at capturing global features, they often include irrelevant information that can detract from the fidelity, potentially compromising the quality of the personalized output, such as including the same background in the generation. To mitigate this issue, some studies incorporate additional prior knowledge to guide the learning process so as to focus on the targeted SoI. For instance, the SoI-specific mask[49, 50, 59, 60, 65, 102] contributes to the effective exclusion of the influence of the background. Moreover, using facial landmarks[6] in the context of human face customization helps improve identity preservation. We will discuss a more detailed technical summary of mask-assistant generation in Section 4.2.

Handling multiple input references presents another challenge but it is essential for real-world deployment. This necessitates an ensemble of features from the multiple reference images to augment the framework's adaptability. Yet, the majority of current PTA systems are limited to supporting one reference input. Some research works[6, 33] propose to average or stack features extracted from multiple references to form a composite SoI representation.

3.2.3 Subject feature fusion

Personalized content synthesis systems typically process two input modalities: Reference images and textual descriptions, as illustrated in Fig. 5. Effective fusion of these heterogeneous features represents a critical technical challenge in PTA frameworks. Current methodologies can be categorized into four primary approaches:

Concatenation-based fusion[5, 9, 33, 34, 62, 95]. This method makes the unique modifier a placeholder token to encapsulate visual subject characteristics. The placeholder token embedding, initialized with image features from visual encoders, is concatenated with text embeddings from the language model. This combined representation subsequently guides generation through standard cross-attention layers in the diffusion process, enabling basic subject-text alignment while maintaining architectural simplicity.

Cross-attention fusion[6, 13, 28, 40, 97, 120, 157]. This paradigm extends the U-Net architecture with specialized attention mechanisms that jointly process visual and textual conditions. For example, IP-Adapter[28] introduces decoupled cross-attention layers that maintain separate query projections for image and text features. In this case, the unique modifier directly displays the plain text.

Multimodal encoder fusion[17, 30, 75, 77, 81]. This approach leverages powerful multimodal encoder architectures (e.g., BLIP-2[195]) to jointly embed visual and textual subject descriptors. BLIP-diffusion[17] exemplifies this strategy by learning a compact subject prompt embedding that fuses image patches with textual names through Q-former modules.

Hybrid fusion[27, 113]. Moreover, some systems integrate multiple fusion strategies. For example, Subject-diffusion[27] integrates both concatenation and cross-attention fusion, taking advantage of the strengths of each approach to enhance the overall personalization capability.

3.2.4 Training data

Training a PTA model for PCS necessitates a large-scale dataset. There are primarily two types of training samples utilized.

Triplet data (reference image, target image, target caption). This dataset format is directly aligned with the PCS objectives, establishing a clear relation between the reference and the personalized content. However, such large-scale triplet samples are not widely available. Several strategies have been proposed to mitigate this issue: 1) Data augmentation. Techniques such as foreground segmentation followed by placement in a different background are used to construct triplet data[17]. 2) Synthetic sample generation. Methods like SuTI[13] utilize multiple TTF models to generate synthetic samples, which are then paired with original references. 3) Utilizing recognizable SoIs. Collecting images of easily recognizable subjects, such as celebrities, significantly facilitates face personalization[23].

Dual data (reference image, reference caption). This dataset is essentially a simplified version of the triplet format, where the personalized content is the original image itself. Such datasets are more accessible, including collections like LAION[196] and LAION-FACE[197]. However, a notable drawback is that training tends to focus more on reconstructing the reference image rather than incorporating the text prompts. Consequently, models trained on this type of data might struggle with complex prompts that require substantial modifications or interactions with objects.

3.3 Hybrid framework

Recently, some works have started to explore the combination of TTF and PTA methods. HyperDreamBooth[26] states that PTA methods provide a general framework that is capable of handling a wide range of common objects, while TTF techniques utilize instance-specific fine-tuning, which can improve the preservation of fine-grained details. They first develop a PTA network, followed by a subject-driven fine-tuning. Similarly, DreamTuner[64] pre-trains a subject encoder that outputs diffusion conditions for accurate reconstruction. Then an additional fine-tuning stage is conducted for fine identity preservation. In contrast, SuTI[13] first applies TTF methods to generate synthetic pairwise samples, which can be used for training the PTA network.

4 Techniques in personalized content synthesis

Building upon the architectural framework discussed in Section 3, this section analyzes learning optimization techniques applicable to both frameworks. We focus on four categories: Attention mechanisms, mask-guided generation, data augmentation and regularization strategies. These methods aim to address key challenges in PCS, like enhancing subject fidelity, minimizing the interference of redundant semantics, enhancing generalization and avoiding overfitting.

4.1 Attention-based operation

Attention-based operations have become a crucial technique in model learning, particularly for processing features effectively[198]. In diffusion models, these operations generally involve manipulating the way a model focuses on different parts of data, often through a method known as the query-key-value (QKV) scheme. While large-scale pre-training has equipped this module with strong feature extraction capabilities, there remains significant ongoing work to enhance its performance for customization tasks.

Explicit attention weight manipulation. A cluster of studies focus on restricting the influence of the SoI token within the attention layers. For example, mix-of-show[20] designs region-aware cross-attention where the feature map is initially generated by the global prompt and replaced with distinct regional features corresponding to each entity. This avoids the misalignment between the words and visual regions. DreamTuner[64] designs a self-subject-attention layer to further refine the subject identity. This attention module takes the features of the generated image as query, the concatenation of generated features as key, and the reference features as value. Layout-control[127] adjusts attention weights specifically around the layout without additional training. Cones 2[21] also defines some negative attention areas to penalize the illegal occupation to allow multiple object generation. VICO[52] inserts a new attention layer where a binary mask is deployed to selectively obscure the attention map between the noisy latent and the reference image features.

Implicit attention guidance. In addition to these explicit attention weights modification methods, many researchers[5, 12, 27, 50, 59, 90, 111] employ localization supervision in the cross-attention module. Specifically, they train cross-attention modules using coordinate-aware loss functions, forcing attention maps to align with annotated subject positions. DreamTuner[64] further refines this approach by designing an attention layer that effectively integrates features from different parts of the image.

4.2 Mask-guided generation

Since reference images contain both the SoI and irrelevant visual elements, masks as a crucial prior indicating the position and contour of the specified object can effectively minimize the influence of redundant information.

Pixel-level mask. Benefitting from advanced segmentation methods like segment anything model (SAM)[199], the SoI can be precisely isolated from the background. Based on this strategy, plenty of studies[49, 50, 59, 60, 65, 90, 99, 102, 110] choose to discard the pixels of the background area so that the reconstruction loss can focus on the targeted object and exclude irrelevant disturbances. Another technique[62] further adds the masked background reconstruction for better disentanglement. In addition, the layout indicated by the pixel mask can be incorporated into the attention modules as a supervision signal[5, 12, 27, 50, 59, 90, 111] to adjust the attention's concentration adaptively. Moreover, the mask can stitch specific feature maps to construct more informative semantic patterns[9, 27, 46].

Feature-level mask. In addition to the pixel-level manipulation, masks can be extended to feature-level operations. DisenBooth[16] defines an identity-irrelevant embedding with a learnable mask. By maximizing the cosine similarity between the identity-preservation embedding and identity-irrelevant embedding, the mask will adaptively exclude the redundant information, and thus the

subject appearance can be better preserved. AnyDoor[116] defines a high-frequency mask that stores detailed SoI features as a condition for the image generation process. Face-diffuser[96] determines the mask through augmentation from the noise predicted by both a pre-trained text-to-image diffusion model and a PTA personalized model. Each model makes its own noise prediction, and the final noise output is a composite created through mask-guided concatenation.

4.3 Data augmentation

Since the optimization of neural networks requires extensive data, existing PCS methods often struggle to capture complete semantic information of the SoI from limited references, resulting in poor images. To address this, various data augmentation strategies are employed to enrich the diversity of the SoI references.

Compositional augmentation. Some methods enhance data diversity through classical image augmentation like blending and spatial rearrangement. SVDiff[12] manually constructs mixed images of multiple SoI as new training data, thereby enhancing the model's exposure to complex scenarios. Such concept composition is also used in other works[27, 152, 151]. BLIP-diffusion[17] segments the foreground subject and composes it in a random background so that the original text-image pairs are expanded to a larger dataset. StyleAdapter[29] chooses to shuffle the image patches to break the irrelevant subject and preserve the desired style. PACGen[110] shows that the spatial position entangles with the identity information. Thus rescaling, center crop and relocation are effective augmentation solutions.

Synthetic data. Generated synthetic data can provide a large amount of training resources when coupled with quality assurance mechanisms. SuTI[13] establishes a cascaded pipeline where TTF models first generate diverse variations of each SoI. These synthetic samples then train the target PTA model. Similarly, DreamIdentity[95] leverages the existing knowledge of celebrities embedded in large-scale pre-trained diffusion model to generate both the source image and the edited face image. StyleDrop[22] and generative active learning (GAL)[39] implement iterative refinement pipelines where high-quality synthetic outputs from early training phases are incorporated into subsequent rounds.

External sources. It is intuitive to leverage web resources to expand training datasets. COTI[47] adopts a scorer network to progressively expand the training set by selecting semantic-relevant samples with high aesthetic quality from a large web-crawled data pool.

4.4 Regularization

Regularization is an effective method that is used to regularize the weight update to avoid overfitting and enhance generalization.

Auxiliary data regularization. To mitigate the overfitting problem that the PCS system consistently produces identical outputs as the references, studies start to use an additional dataset composed of images with the same category of the SoI[4]. By reconstructing these images, the personalized model is required to generate diverse instances of the class while adapting to the target subject. Building on this strategy, StyleBoost[83] introduces an auxiliary style-specific data to separate content and aesthetic adaptation. Later, a dataset[55] is curated with detailed textual prompts (specifying attributes/contexts) to improve disentanglement between subject characteristics and background features.

Text embedding constraints. The semantic richness of pre-trained text (e.g., subject class name) provides a powerful regularization signal for personalized generation. By strategically constraining how subject-specific representations interact with established linguistic concepts in the embedding space, these approaches can achieve better generalization ability. For example, perfusion[15] constrains key projections toward class noun embeddings while learning value projections from subject images. Inspired by coached active learning[200, 201], which uses anchor concepts for optimization guidance, compositional inversion[35] employs a set of semantically related tokens as anchors to constrain the token embedding search. In addition, some works[65, 74] regularize learnable token offsets relative to pre-trained CLIP embeddings. By minimizing the offset, the final word embedding is able to achieve better text alignment. Similarly, Cones 2[21] minimizes the offset by reconstructing the features of 1 000 sentences containing the class noun. And [37] optimizes the learnable token towards the mean textual embedding of 691 well-known names. Domain-agnostic[25] proposes to use a contrastive loss to guide the SoI text embedding close to its nearest CLIP tokens pre-trained on large-scale samples. On the other hand, VICO[52] empirically finds that the end-of-text token ⟨|EOT|⟩ keeps the semantic consistency of SoI. To leverage this discovery, an L2 loss is leveraged to reduce the difference of attention similarity logits between the SoI token and ⟨|EOT|⟩.

5 Categorization of image personalization tasks

As shown in Fig. 6, personalization covers a range of areas, including objects, styles, faces, etc. This section analyzes these tasks through the frameworks established in Section 3, examining both TTF and PTA approaches for each domain. We also summarize these studies in Tables 1–3 to provide a clear overview and facilitate quick comparison.

5.1 Personalized object generation

As the foundational task, personalized object genera-

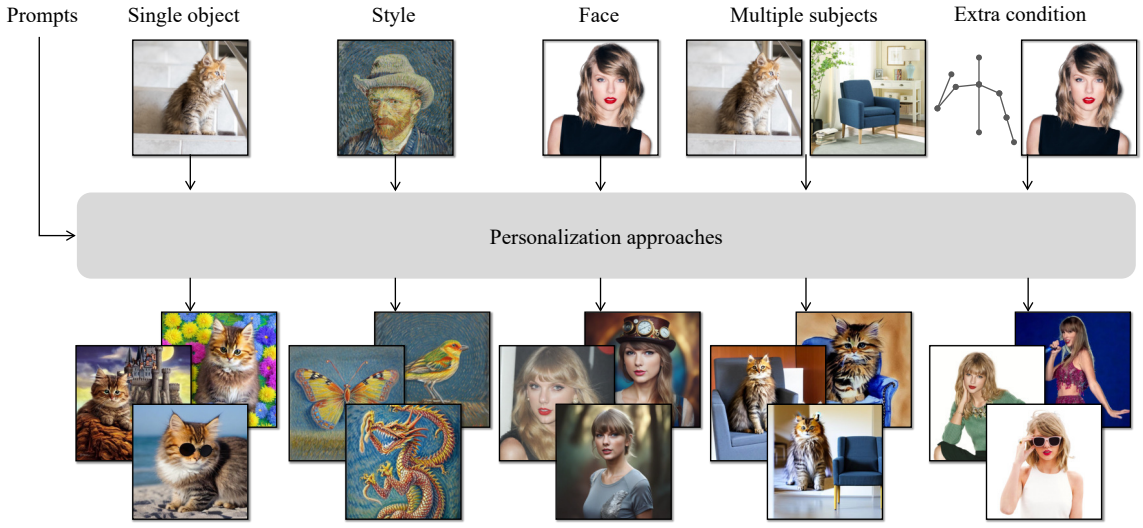

Fig. 6 Visualization of several personalized image synthesis tasks. Given text prompts and a few images of the subject of interest, personalization approaches are required to produce expected images.

tion requires learning discriminative features from general instances (e.g., toys, vehicles, or architecture) and rendering them in novel contexts specified by textual prompts.

TTF framework. TTF methods perform instance-specific optimization during inference by fine-tuning model parameters on reference images. This achieves superior subject fidelity and handles rare attributes effectively.

An important branch in TTF methods is to optimize the learnable token embeddings. The first work starts from textual inversion[7], which applies a simple yet effective method that introduces the unique modifier as a new token to represent the SoI. One of the significant benefits of this method is its minimal storage requirement as the new tokens consume just a few kilobytes. However, the method compresses complex visual features into a small set of parameters, which can lead to long convergence times and a potential loss in visual fidelity. Recent work aims to address these limitations through several directions. DVAR[43] improves training efficiency by proposing a clear stopping criterion by removing all randomness to indicate the convergence. For enhanced representation, P+[11] introduces distinct learnable tokens across different layers of the U-Net architecture, thereby offering better attribute control through additional learnable parameters. NeTI[18] advances this concept by proposing a neural mapper that adaptively outputs token embeddings based on the denoising timestep and specific U-Net layers. ProSpect[19] recommends optimizing multiple token embeddings tailored to different denoising timesteps based on the observation that different types of prompts, like layout, color, structure and texture, are activated at different stages of the denoising process. Similarly, a study by [57] shows layered activation insight to learn distinct attributes by selectively activating the tokens within their respective scopes. Later, HiFiTuner[60] integrates multiple techniques into the learnable token, including mask-guided loss function, parameter regularization, time-dependent embedding and generation refinement assisted by the nearest reference. Alternative approaches, like DreamArtist[42], choose to optimize both negative and positive prompt embeddings to refine the detail preservation. In addition to these token-level refinement approaches, the field continues to evolve with novel techniques such as InstructBooth's reinforcement learning framework[61] and gradient-free evolutionary optimization[48]. In summary, recent developments following the foundational work of textual inversion focus on reducing training times and enhancing the visual quality of generated images.

In the realm of TTF methods for PCS, there is a clear shift towards fine-tuning model weights rather than just the token embeddings. This approach often addresses the limitations where token embeddings alone struggle to capture complex semantics uncovered in the pre-training data[20, 58]. DreamBooth[4] proposes to use a unique modifier by a rare token to represent the SoI and fine-tune the whole parameters of the diffusion model. Besides, a regularization dataset containing 20–30 images with the same category as SoI is adopted to overcome the overfitting problem. These two combined approaches achieve impressive performance that largely promotes the progress of the research on image personalization. However, fine-tuning the entire model for each new object causes considerable storage costs, potentially rendering widespread application. To address this, custom diffusion[8] focuses on identifying and fine-tuning critical parameters, particularly the key-value projections in cross-attention layers, to achieve a balance of visual fidelity and storage efficiency. Further approach, perfusion[15], also adopts the cross-attention fine-tuning and proposes to regularize the update direction of the K (key) projection towards the super-category token embedding and the V (value) projection towards the learnable token embedding. COMCAT[51] introduces a low-rank approximation of atten-

tion matrices, which drastically reduces storage requirements to 6 MB while maintaining high fidelity in the outputs. Additionally, methods like adapters[22, 45] and LoRA variants[20, 26, 38, 54, 65, 102] are increasingly utilized in personalized generation for parameter-efficient fine-tuning. It is worth noting that the token embedding fine-tuning is compatible with diffusion weight fine-tuning. Multiple methods[50, 64, 90] have started using combined weight fine-tuning.

PTA framework. For practical deployment of PCS systems, fast response time is a crucial factor. PTA methods enable real-time generation (less than 10 sec/subject) by leveraging large-scale pre-training to avoid per-subject optimization during the inference phase. Re-Imagen[72] introduces a retrieval-augmented generative approach, which leverages features from text-image pairs retrieved via a specific prompt. While it is not specifically tailored for object personalization, it demonstrates the feasibility of training reference-conditioned frameworks. Later, ELITE[9] specifically targets image personalization by combining the global reference features with text embedding while incorporating local features that exclude irrelevant backgrounds. Both fused features and local features serve as conditions for the denoising process. Similarly, InstantBooth[14] retrains CLIP models to extract image features and patch features, which are injected into the diffusion model via the attention mechanism and learnable adapter, respectively. Additionally, UMM-diffusion[75] designs a multi-modal encoder that produces fused features based on the reference image and text prompt. The text features and multi-modal hidden state are seen as guidance signals to predict a mixed noise. Another work, SuTI[13], adopts the same architecture as Re-Imagen. The difference lies in the training samples which are produced by a massive number of TTF models, each is tuned on a particular subject set. This strategy promotes a more precise alignment with personalization at an instance level rather than the class level of Re-Imagen. Moreover, Domain-agnostic[25] combines a contrastive-based regularization technique to push the pseudo embedding produced by the image encoder towards the existing nearest pre-trained token. Besides, they introduce a dual-path attention module separately conditioned on the nearest token and pseudo embedding. Compared to the methods that use separate encoders to process a single modality, some works have explored the usage of pre-trained multi-modal large language models (MLLM) that can process text and image modality within a unified framework. For example, BLIP-diffusion[17] utilizes the pre-trained BLIP2[195] that encodes multimodal inputs including the SoI reference and a class noun. The output embedding is then concatenated with context description and serves as a condition to generate images. Further, customization assistant[77] and KOSMOS-G[30] replace the text encoder of stable diffusion with a pre-trained MLLM to output a fused feature based on the reference and context description. Meanwhile, to meet the standard format of stable diffusion, a network is trained to align the dimension of the output embedding.

5.2 Personalized style generation

Personalized style generation seeks to tailor the aesthetic elements of reference images. The concept of "style" now includes a wide range of artistic elements, such as brush strokes, material textures, color schemes, structural forms, lighting techniques and cultural influences.

TTF framework. In this field, StyleDrop[22] leverages adapter tuning to efficiently capture the style from a single reference image. This method demonstrates the effectiveness through iterative training, utilizing synthesized images refined by feedback mechanisms, like human evaluations and CLIP scores. This approach not only enhances style learning but also ensures that the generated styles align closely with human aesthetic judgments. Later, GAL[39] proposes an uncertainty-based evaluation strategy to filter high-quality synthetic-style data and uses a weighted schema to balance the contribution of the additional samples and the original reference. Furthermore, StyleAligned[84] focuses on maintaining stylistic consistency across a batch of images. This is achieved by using the first image as a reference, which acts as an additional key and value in the self-attention layers, ensuring that all subsequent images in the batch adhere to the same stylistic guidelines. Style-friendly[86] introduces a novel diffusion model fine-tuning approach that enhances personalized artistic style generation by adaptively biasing noise sampling toward higher noise levels, where stylistic features emerge.

PTA framework. For the PTA framework, StyleAdapter[29] employs a dual-path cross-attention mechanism within the PTA framework. This model introduces a specialized embedding module designed to extract and integrate global features from multiple style references. Diptych prompting[82] utilizes an inpainting mechanism to draw another image with the same style of the reference part.

5.3 Personalized face generation

Personalized face generation aims to generate diverse identity images that adhere to text prompt specifications, utilizing only a few initial face images. Compared to general object personalization, the scope is narrowed to a specific class, and humans. An obvious benefit is that one can readily leverage large-scale human-centric datasets[197, 202, 203] and utilize pre-trained models in well-developed areas, like face landmark detection[204] and face recognition[205].

TTF framework. Regarding TTF methods, PromptNet[88] trains a diffusion-based network that encodes the

input image and noisy latent features to a word embedding. To alleviate the overfitting problem, the noises predicted by the word embedding and context description are balanced through fusion sampling in classifier-free guidance. Additionally, Celeb basis[23] provides a novel idea that the personalized ID can be viewed as the composition of celebrity's face, which has been learned by the pre-trained diffusion model. Based on this hypothesis, a simple multi-layer perceptron (MLP) is optimized at test time to transform face features into the weighting of different celebrity name embeddings.

PTA framework. Thanks to the abundance of available datasets featuring the same individual in various contexts, which provide valuable data for pre-training PTA methods, the number of works in PTA frameworks is rapidly increasing. Face0[24] crops the face region to extract refined embeddings and concatenates them with text features. During the sampling phase, the output of classifier-free guidance is replaced by a weighted combination of the noise patterns predicted by face-only embedding, text-only embedding and concatenated face-text embedding. The $\mathcal{W}+$ adapter[97] constructs a mapping network and residual cross-attention modules to transform the facial features from the StyleGAN[206] $\mathcal{W}+$ space into the text embedding space of stable diffusion. FaceStudio[32] adapts the cross-attention layer to support hybrid guidance including stylized images, facial images and textual prompts. Moreover, PhotoMaker[33] constructs a high-quality dataset through a meticulous data collection and filtering pipeline. They use a two-layer MLP to fuse ID features and class embeddings for an overall representation of human portrait. PortraitBooth[34] also employs a simple MLP, which fuses the text condition and shallow features of a pre-trained face recognition model. To ensure expression manipulation and facial fidelity, they add another expression token and incorporate the identity preservation loss and mask-based cross-attention loss. InstantID[6] additionally introduces a variant of ControlNet that takes facial landmarks as input, providing stronger guiding signals compared to the methods that solely rely on attention fusion.

5.4 Multiple subject composition

Multiple subject composition refers to the scenario where users intend to compose multiple SoI displayed in one or more references.

TTF framework. This task presents a challenge for the TTF methods, particularly in how to integrate the parameters which are separately fine-tuned for individual SoI within the same module. Some works focus on the one-for-one generation, following a fusion mechanism. For instance, Custom diffusion[8] proposes a constrained optimization method to merge the cross-attention key-value projection weights with the goal of maximizing reconstruction performance for each subject. Mix-of-show[20] fuses the LoRA[207] weights with the same optimization objective. StyleDrop[22] dynamically summarizes noise predictions from each personalized diffusion model. In OMG[38], the latent features predicted by each LoRA-tuned model is spatially composited using the subject mask. Joint training is another strategy to cover all expected subjects. SVDiff[12] employs a data augmentation method called cut-mix to compose several subjects together and applies a location loss to regularize attention maps, ensuring alignment between each subject and its corresponding token. Similar strategies are found in other works[8, 23] which train a single model by reconstructing the appearance of every SoI. There are advanced control mechanisms designed to manage multiple subjects. Cones[10] proposes to find a small cluster of neurons that preserve the most information about SoI. The neurons belonging to different SoI will be simultaneously activated to generate the combination. Compositional inversion[35] introduces spatial region assignment to different subjects to improve the composition success rate.

PTA framework. For PTA frameworks, multi-subject generation is achieved through specialized architectural designs. Fastcomposer[5], subject-diffusion[27], and λ-ECLIPSE[119] place each subject feature in its corresponding placeholder in the text embedding, ensuring a seamless and efficient combination. CustomNet[117] and MIGC[118] train a PTA network that supports location control for each subject. SSR-encoder[40] implements an encoder to selectively preserve the desired subject feature and a cross-attention module to support multi-subject feature fusion.

5.5 High-level semantic personalization

Recently, the field of image personalization has started to include complex semantic relationships and high-level concepts. Different approaches have been developed to enhance the capability of models to understand and manipulate these abstract elements.

TTF framework. For now, all researches in this field are based on the TTF framework. ReVersion[121] intends to invert object relations from references. Specifically, they use a contrastive loss to guide the optimization of the token embedding towards specific clusters of part-of-speech tags, such as prepositions, nouns and verbs. Meanwhile, they also increase the likelihood of adding noise at the larger timesteps during the training process to emphasize the extraction of high-level semantic features. Lego[31] focuses on more general concepts, such as adjectives, which are frequently intertwined with the subject appearance. The concept can be learned from a contrastive loss applied to the dataset comprising clean subject images and images that embody the desired adjectives. Moreover, ADI[124] aims to learn the action-specific identifier from the references. To ensure that the inversion only focuses on the desired action, ADI extracts gradient in-

variance from a constructed triplet sample and applies a threshold to mask out the irrelevant feature channels.

6 Extensions of personalized content synthesis

While the core PCS system focuses on image generation from reference subjects, recent advances have expanded its capabilities across multiple dimensions. This section examines several frontier extensions that push the boundaries of personalization technology.

6.1 Personalization on extra conditions

Recent personalization tasks tend to include additional conditions for diverse content customization. One common application is to customize the subject into a fixed source image. For example, PhotoSwap[125] introduces a new task that replaces the subject in the source image with the SoI from the reference image. To address this requirement, they first fine-tune a diffusion model on the references to obtain a personalized model. To preserve the background of the source image, they initialize the noise with denoising diffusion implicit models (DDIM) inversion[208] and replace the intermediate feature maps with those derived from source image generation during inference time. Later, MagiCapture[97] broadens the scope to face customization. Another similar application can be found in virtual try-on, which aims to fit selected clothing onto a target person. The complexities of this task have been thoroughly analyzed in another survey[209].

Additional conditions in personalization tasks may include adjusting the layout[127], transforming sketches[101], controlling viewpoint[117, 129], or modifying poses[6]. Each of these conditions presents unique challenges and requires specialized approaches to integrate these elements seamlessly into the personalized content.

6.2 Personalized video generation

With the rising popularity of video generation[210], video personalization has also begun to attract attention. In video personalization, the SoI can be classified into three distinct categories: Appearance, motion and the combination of both appearance and motion.

Appearance-based video personalization. This task focuses on transferring subject appearance from static images to video sequences. The standard TTF pipeline utilizes reference images as appearance anchors and fine-tunes video diffusion models (VDM) for temporal synthesis. The process involves leveraging sophisticated methods from 2D personalization, such as parameter-efficient fine-tuning[151], data augmentation[152, 151] and attention manipulation[154, 157, 151]. Additionally, several studies[153, 154, 157, 161] have explored the PTA framework. These diffusion models are specifically tailored to synthesize videos based on the image references.

Motion-based video personalization. In this task, the reference input switches to a video clip containing a consistent action. A common approach is to fine-tune the video diffusion model by reconstructing the action clip[136, 137, 139, 141–144]. However, distinguishing between appearance and motion within the reference video can be challenging. To solve this problem, SAVE[143] applies appearance learning to ensure that appearance is excluded from the motion learning phase. Additionally, VMC[142] removes the background information during training prompt construction.

Appearance and motion personalization. When integrating both subject appearance and motion, innovative methods are employed to address the complexities of learning both aspects simultaneously. MotionDirector[140] utilizes spatial and temporal losses to facilitate learning across these dimensions. Another approach, DreamVideo[145], incorporates residual features from a randomly selected frame to emphasize subject information. This technique enables the fine-tuned module to primarily focus on learning motion dynamics.

In summary, video personalization strategies vary significantly based on the specific aspects. Moreover, due to the current limitations in robust video feature representation, PTA video personalization that is directly conditioned on video input remains an area under exploration.

6.3 Personalized 3D generation

Personalized 3D generation refers to the process of creating customized 3D models or scenes based on 2D SoI images. Basically, the pipeline begins by fine-tuning a 2D diffusion model using TTF methods. This tuned model then utilizes score distillation sampling (SDS)[211] to train a 3D neural radiance field (NeRF) model[212] for each specific prompt[162, 167]. Building on this foundation, several methods have been developed to improve the workflow. DreamBooth3D[163] structures the process into three phases: Initializing and optimizing an NeRF from a DreamBooth model, rendering multi-view images, and fine-tuning a secondary DreamBooth for the final 3D NeRF refinement. Consist3D[168] enhances text embeddings by training two distinct tokens, a semantic and a geometric token, during 3D model optimization. TextureDreamer[171] focuses on extracting texture maps from optimized spatially-varying bidirectional reflectance distribution fields (BRDF) for rendering texture on wide-range 3D subjects.

Additionally, advancements extend to 3D avatar rendering and dynamic scenes. Animate124[169] and Dream-in-4D[170] integrate video diffusion for 4D dynamic scene support within the 3D optimization process. In avatar rendering, PAS[164] generates 3D body poses configurable by avatar settings, StyleAvatar3D[165] facilitates 3D

avatar generation based on images, and AvatarBooth[166] employs dual fine-tuned diffusion models for separate face and body generation.

6.4 Attack and defense

This rapid advancement raises concerns about the ethical implications of PCS, particularly in areas such as misinformation, privacy violations and the creation of deepfakes. There is an increased risk that individuals or organizations may exploit them to produce misleading content or manipulate public perception. To mitigate this, Anti-DreamBooth[173] aims to add a subtle noise perturbation to the references so that any personalized model trained on these samples only produces terrible results. The basic idea is to maximize the reconstruction loss of the surrogate model. Additionally, Wu et al.[174] suggest to predefine a collection of trigger words and meaningless images. These data are paired and incorporated during the training phase. Once the trigger words are encountered, the synthesized image will be intentionally altered for safeguarding.

6.5 Other emerging directions

Several works are exploring different personalization extensions. For example, scalable vector graphics (SVG) personalization is introduced by [177], in which a parameter-efficient fine-tuning method is applied to create SVGs. After first-step generation, the SVGs are refined through a process that includes semantic alignment and a dual optimization approach, which utilizes both image-level and vector-level losses to enhance the final output. Another application, 360-degree panorama customization[178], is also emerging as a potential tool for personalization in the digital imaging realm.

7 Evaluation

7.1 Evaluation dataset

To assess the performance of personalized models, various datasets have been developed:

DreamBench[4] serves as the primary evaluation benchmark for DreamBooth[4], containing 30 diverse subjects (e.g., backpacks, animals, vehicles and toys) with 25 unique prompts per subject.

DreamBench-v2[13] expands the evaluation scope of DreamBench by adding 220 test prompts for each subject.

Custom-10[8] used in custom diffusion[8] evaluates 10 subjects, each with 20 specific test prompts, and includes tests for multi-subject composition with 5 pairs of subjects and 8 prompts for each pair.

Custom-101[8] is the latest released dataset by the authors of custom diffusion[8], which comprises 101 subjects to provide a broader scope of evaluation.

Stellar[102] specifically targets human-centric evaluation, featuring 20 000 prompts on 400 human identities.

Despite these contributions, they remain fragmented across different research groups and the research community still lacks a benchmark tested on a large number of personalized generation tasks. To address this gap, this survey introduces a comprehensive evaluation dataset[1] designed for the most common personalized object and face personalization. Section 7.3 details our benchmark design and evaluation results.

7.2 Evaluation metrics

As PCS aims to maintain fidelity to the SoI while ensuring alignment with textual conditions, the metrics are designed from two aspects, text alignment and visual fidelity.

The text alignment metrics quantify how precisely generated outputs reflect prompt semantics:

CLIP-T measures semantic alignment using the cosine similarity between CLIP[193] embeddings of generated images and their text prompts.

ImageReward[213], **HPS score** (v1/v2)[214, 215], and **PickScore**[216] employ learned models trained on human judgments to better correlate with perceptual quality to reflect human preference better.

To determine how closely the generated subject resembles the SoI, visual fidelity can be assessed via the following metrics:

CLIP-I evaluates subject preservation through CLIP image embedding similarity between generations and references. Optimal values should balance fidelity (high scores) against overfitting (excessive scores that ignore text guidance).

DINO-I[217] provides a complementary assessment using DINO's instance-aware features, particularly effective for object-level similarity.

Fréchet inception distance (FID)[218] quantifies the statistical similarity between generated and real image distributions through Inception-V3[219].

In addition to these commonly adopted metrics, there are some discussions of specialized metrics for PCS system evaluation: LyCORIS[54] introduces a 5-dimensional assessment covering fidelity, controllability, diversity, base model preservation and image quality. Stellar[102] develops six human-centric metrics including soft-penalized CLIP text score, identity preservation score, attribute preservation score, stability of identity score, grounding objects accuracy, and relation fidelity score. These metrics ensure a structured and detailed evaluation of personalized models. This evolving metric landscape reflects the

[1] This dataset is available on: https://github.com/zhangxulu1996/awesome-personalization

growing sophistication of PCS systems, with optimal evaluation typically requiring combinations of general and task-specific measures.

7.3 New benchmark on SoTA methods

Although multiple evaluation datasets have been proposed in this area, there is still an urgent need for a standardized benchmark that systematically assesses performance across diverse PCS methodologies. To address this limitation, we present a new test dataset called persona with a comprehensive evaluation built upon existing works.

For object. Persona includes 47 subjects from available resources[4, 7, 8]. Following the methodology of DreamBooth[4], we categorize the subjects into two classes: Objects and live pets, based on whether the subject is a living entity. Specifically, 10 out of 47 subjects are pets and the remaining 37 are various objects. To evaluate the performance, we utilize the text prompts from DreamBooth[4]. This includes 20 recontextualization prompts and 5 property modification prompts for objects, along with 10 recontextualization, 10 accessorization and 5 property modification prompts for pets, totally 25 prompts per category.

For face. We also collected 15 subjects from CelebA[220] into our Persona dataset. We use 40 prompts for evaluation, including 10 accessory prompts, 10 style prompts, 10 action prompts and 10 context prompts.

Settings. To assess existing representative PCS methods using our constructed test dataset, we compute CLIP-T to assess the text alignment. We calculate CLIP-I for subject fidelity in object generation. For subject fidelity in face generation, we detect faces in both the generated images and target images using multi-task cascaded convolutional neural network (MTCNN)[221] and calculate the pairwise identity similarity using FaceNet[222]. To uniformly evaluate and compare existing personalized generation methods, we select 22 representative PCS methods for evaluation. We generate 4 images for each test prompt and set the same random seed for all methods.

Results. The evaluation results are shown in Table 4. It is evident that no method excels simultaneously in both visual fidelity and text alignment metrics. This highlights a significant challenge currently faced by PCS methods: Achieving the optimal trade-off between subject preservation and editability. Striking this balance is difficult, as high subject fidelity often comes at the cost of prompt fidelity and vice versa. Furthermore, we note that higher visual fidelity does not always equate to better performance. The generated images sometimes exhibit patterns that closely mirror the reference images and ignore the prompt guidance. This phenomenon primarily arises from model overfitting on the reference input, which hinders the model's ability to generalize. Consequently, the visual similarity metric yields higher results based on the mirrored output and the reference, rather than providing an accurate representation of the performance.

Table 4 Evaluation results of representative methods on our persona evaluation dataset

Type	Methods	Framework	Backbone	CLIP-T	CLIP-I
Object	Textual inversion[7]	TTF	SD 1.5	0.199	0.749
	DreamBooth[4]	TTF	SD 1.5	0.286	0.772
	P+[11]	TTF	SD 1.4	0.244	0.643
	Custom diffusion[8]	TTF	SD 1.4	0.307	0.722
	NeTI[18]	TTF	SD 1.4	0.283	0.801
	SVDiff[12]	TTF	SD 1.5	0.282	0.776
	Perfusion[15]	TTF	SD 1.5	0.273	0.691
	ELITE[9]	PTA	SD 1.4	0.292	0.765
	BLIP-Diffusion[17]	PTA	SD 1.5	0.292	0.772
	IP-Adapter[28]	PTA	SD 1.5	0.272	**0.825**
	SSR encoder[40]	PTA	SD 1.5	0.288	0.792
	MoMA[81]	PTA	SD 1.5	0.322	0.748
	Diptych prompting[82]	PTA	FLUX 1.0 dev	**0.327**	0.722
	λ-ECLIPSE[119]	PTA	Kandinsky 2.2	0.272	0.824
	MS-diffusion[120]	PTA	SDXL	0.298	0.777
Face	Cross initialization[37]	TTF	SD 2.1	0.261	0.469
	Face2Diffusion[107]	PTA	SD 1.4	0.265	0.588
	SSR encoder[40]	PTA	SD 1.5	0.233	0.490
	FastComposer[5]	PTA	SD 1.5	0.230	0.516
	IP-Adapter[28]	PTA	SD 1.5	0.292	0.462
	IP-Adapter[28]	PTA	SDXL	0.292	0.642
	PhotoMaker[33]	PTA	SDXL	**0.311**	0.547
	InstantID[6]	PTA	SDXL	0.278	**0.707**

8 Challenges and outlook

8.1 Overfitting problem

As discussed in Section 7.3, current PCS systems face a critical challenge of overfitting because of a limited set of reference images. This overfitting problem manifests in two ways: 1) Loss of SoI editability. The personalized model tends to produce images that rigidly mirror the SoI in the reference, such as consistently depicting a cat in an identical pose. 2) Irrelevant semantic inclusion. The irrelevant elements in the references are generated in the output, such as backgrounds or objects that are not pertinent to the current context.

To investigate the rationale behind, compositional inversion[35] observes that the learned token embedding is

located in an out-of-distribution area compared to the center distribution formed by pre-trained words. This is also found in another work[37] that the learnable token embeddings deviate significantly from the distribution of the initial embedding. In addition, there is evidence[35, 46, 52] suggesting that the unique modifier dominates in the cross-attention layers compared to the other context tokens, leading to the absence of other semantic appearances.

To address this issue, many solutions have been proposed. Most methods discussed in Section 4 contribute to the alleviation of the overfitting problem, such as the exclusion of redundant background, attention manipulation, regularization of the learnable parameters and data augmentation. However, it has not been solved yet, especially in the cases where the SoI has a non-rigid appearance[35] or the context prompt has a similar semantic correlation with the irrelevant elements in the reference[39]. It is clear that addressing overfitting in PCS is not merely a technical challenge but a necessity for ensuring the practical deployment and scalability of these systems in varied and dynamic real-world environments. Therefore, an effective strategy and robust evaluation metrics are urgently needed to achieve broader adoption and greater satisfaction in practical applications.

8.2 Trade-off on subject fidelity and text alignment

The ultimate goal of personalized content synthesis is to create systems that not only render the SoI with high fidelity but also effectively respond to textual prompts. However, achieving excellence in both areas simultaneously presents a notable conflict. Specifically, high subject fidelity typically involves capturing and reproducing detailed and specific features of the SoI. This often requires the model to minimize the reconstruction loss to replicate delicate characteristics accurately. Conversely, text alignment necessitates that the system flexibly adapts the SoI according to varying textual descriptions. These descriptions may suggest changes in pose, expression, environment or stylistic alterations that do not aim to reconstruct the exact visualization in the reference. As a result, it becomes challenging to achieve flexible adaptation in different contexts while simultaneously pushing the model to capture fine-grained details. To address this inherent conflict, perfusion[15] proposes to regularize the attention projections by these two items. Cao et al.[91] decouple the conditional guidance into two separate processes, which allows for the distinct handling of subject fidelity and textual alignment. Despite these efforts, there still remains room for further exploration and refinement of this issue. Enhanced model architectures, innovative training methodologies and more dynamic data handling strategies could potentially provide new pathways to better balance the demands of subject and text fidelity in PCS systems.

8.3 Standardization and evaluation

Despite the popularity of personalization, there is a noticeable lack of standardized test datasets and robust evaluation metrics that accurately capture the performance of different strategies. Currently, a widely used metric for assessing visual fidelity relies on CLIP image similarity. However, this approach may wrongly exaggerate the value when the model is overfitted to the references. Therefore, future efforts should focus on creating comprehensive and widely accepted benchmarks that can evaluate various aspects of PCS models, including but not limited to visual fidelity and subject editability.

8.4 Multimodal autoregressive frameworks

Recent advancements in multimodal autoregressive models present novel solutions for PCS by unifying cross-modal understanding and generation. Models like Emu3[223] demonstrate that autoregressive architectures can natively handle image-text-video sequences through discrete tokenization and joint transformer training. This paradigm enables seamless integration of user-provided multimodal references (e.g., text descriptions paired with SoI images) while maintaining contextual coherence across generation steps. Besides, this framework natively supports for subject editing via multi-round chat, effectively addressing the overfitting limitations commonly observed in diffusion-based models.

8.5 Interactive personalization workflow

The evolution of interactive generation systems has unlocked new frontiers for PCS, particularly through the integration of multi-round interactive generation. This capability allows users to iteratively refine and accurately define the SoI, addressing the challenge of translating vague or complex requirements into precise content generation. For instance, conversational PCS systems like Gemini-2.0-flash[224] exemplify this progress, leveraging natural language dialogue to iteratively optimize both subject fidelity and prompt alignment. By enabling users to provide real-time feedback and adjust parameters through chat-like interactions, these systems bridge the gap between abstract intent and concrete outputs, aligning with PCS's core objective of balancing faithful subject representation with flexible editability.

9 Conclusions

This survey has provided a thorough review of personalized content synthesis with diffusion models, particularly focusing on 2D image customization. We explore two main frameworks, TTF and PTA methods, and delve into their mechanics. We also cover the recent progress in

specific customization areas, including object, face, style, video and 3D synthesis. In addition to the impressive techniques, we propose several challenges that still need to be addressed. These challenges include preventing overfitting, finding the right balance between reconstruction quality and editability, and standardizing evaluation methods. To support ongoing research, we collect a test dataset from existing literature and evaluate the classical method to provide a clear comparison. By providing detailed analysis and outlining targeted recommendations, we hope to promote further innovation and collaboration within the PCS community.

Acknowledgements

This work was supported in part by Chinese National Natural Science Foundation Projects, China (Nos. U23B2054, 62276254 and 62372314), Beijing Natural Science Foundation, China (No. L221013), InnoHK program, and Hong Kong Research Grants Council through Research Impact Fund, China (No. R1015-23). Open access funding provided by The Hong Kong Polytechnic University, China.

Declarations of conflict of interest

The authors declared that they have no conflicts of interest to this work.

Open Access

References

[1] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q. L. Han, Y. Tang. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, 2023. DOI: 10.1109/JAS.2023.123618.

[2] F. A. Croitoru, V. Hondru, R. T. Ionescu, M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10850–10869, 2023. DOI: 10.1109/TPAMI.2023.3261988.

[3] V. Uc-Cetina, N. Navarro-Guerrero, A. Martin-Gonzalez, C. Weber, S. Wermter. Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1543–1575, 2023. DOI: 10.1007/s10462-022-10205-5.

[4] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 22500–22510, 2023. DOI: 10.1109/CVPR52729.2023.02155.

[5] G. Xiao, T. Yin, W. T. Freeman, F. Durand, S. Han. FastComposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, vol. 133, no. 3, pp. 1175–1194, 2025. DOI: 10.1007/s11263-024-02227-z.

[6] Q. Wang, X. Bai, H. Wang, Z. Qin, A. Chen, H. Li, X. Tang, Y. Hu. InstantID: Zero-shot identity-preserving generation in seconds, [Online], Available: https://arxiv.org/abs/2401.07519, 2024.

[7] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[8] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, J. Y. Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 1931–1941, 2023. DOI: 10.1109/CVPR52729.2023.00192.

[9] Y. Wei, Y. Zhang, Z. Ji, J. Bai, L. Zhang, W. Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 15897–15907, 2023. DOI: 10.1109/ICCV51070.2023.01461.

[10] Z. Liu, R. Feng, K. Zhu, Y. Zhang, K. Zheng, Y. Liu, D. Zhao, J. Zhou, Y. Cao. Cones: Concept neurons in diffusion models for customized generation. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, Article number 890, 2023.

[11] A. Voynov, Q. Chu, D. Cohen-Or, K. Aberman. P+: Extended textual conditioning in text-to-image generation, [Online], Available: https://arxiv.org/abs/2303.09522, 2023.

[12] L. Han, Y. Li, H. Zhang, P. Milanfar, D. Metaxas, F. Yang. SVDiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 7289–7300, 2023. DOI: 10.1109/ICCV51070.2023.00673.

[13] W. Chen, H. Hu, Y. Li, N. Ruiz, X. Jia, M. W. Chang, W. W. Cohen. Subject-driven text-to-image generation via apprenticeship learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 30286–30305, 2023.

[14] J. Shi, W. Xiong, Z. Lin, H. J. Jung. InstantBooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, Seattle, USA, pp. 8543–8552, 2024. DOI: 10.1109/CVPR52733.2024.00816.

[15] Y. Tewel, R. Gal, G. Chechik, Y. Atzmon. Key-locked rank one editing for text-to-image personalization. In *Proceedings of ACM SIGGRAPH Conference*, Los Angeles, USA, Article number 12, 2023. DOI: 10.1145/3588432.3591506.

[16] H. Chen, Y. Zhang, S. Wu, X. Wang, X. Duan, Y. Zhou, W. Zhu. DisenBooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[17] D. Li, J. Li, S. C. H. Hoi. BLIP-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 30146–30166, 2023.

[18] Y. Alaluf, E. Richardson, G. Metzer, D. Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM Transactions on Graphics*, vol. 42, no. 6, Article number 243, 2023. DOI: 10.1145/3618322.

[19] Y. Zhang, W. Dong, F. Tang, N. Huang, H. Huang, C. Ma, T. Y. Lee, O. Deussen, C. Xu. ProSpect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics*, vol. 42, no. 6, Article number 244, 2023. DOI: 10.1145/3618342.

[20] Y. Gu, X. Wang, J. Z. Wu, Y. Shi, Y. Chen, Z. Fan, W. Xiao, R. Zhao, S. Chang, W. Wu, Y. Ge, Y. Shan, M. Z. Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing System*, New Orleans, USA, Article number 699, 2023.

[21] Z. Liu, Y. Zhang, Y. Shen, K. Zheng, K. Zhu, R. Feng, Y. Liu, D. Zhao, J. Zhou, Y. Cao. Cones 2: Customizable image synthesis with multiple subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 2508, 2023.

[22] K. Sohn, N. Ruiz, K. Lee, D. C. Chin, I. Blok, H. Chang, J. Barber, L. Jiang, G. Entis, Y. Li, Y. Hao, I. Essa, M. Rubinstein, D. Krishnan. StyleDrop: Text-to-image generation in any style. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 2920, 2023.

[23] G. Yuan, X. Cun, Y. Zhang, M. Li, C. Qi, X. Wang, Y. Shan, H. Zheng. Inserting anybody in diffusion models via celeb basis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 3190, 2023.

[24] D. Valevski, D. Lumen, Y. Matias, Y. Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *Proceedings of SIGGRAPH Asia Conference Papers*, Sydney, Australia, Article number 94, 2023. DOI: 10.1145/3610548.3618249.

[25] M. Arar, R. Gal, Y. Atzmon, G. Chechik, D. Cohen-Or, A. Shamir, A. H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *Proceedings of SIGGRAPH Asia Conference Papers*, Sydney, Australia, Article number 72, 2023. DOI: 10.1145/3610548.3618173.

[26] N. Ruiz, Y. Li, V. Jampani, W. Wei, T. Hou, Y. Pritch, N. Wadhwa, M. Rubinstein, K. Aberman. HyperDreamBooth: HyperNetworks for fast personalization of text-to-image models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6527–6536, 2024. DOI: 10.1109/CVPR52733.2024.00624.

[27] J. Ma, J. Liang, C. Chen, H. Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *Proceedings of ACM SIGGRAPH Conference Papers*, Denver, USA, Article number 25, 2024. DOI: 10.1145/3641519.3657469.

[28] H. Ye, J. Zhang, S. Liu, X. Han, W. Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models, [Online], Available: https://arxiv.org/abs/2308.06721, 2023.

[29] Z. Wang, X. Wang, L. Xie, Z. Qi, Y. Shan, W. Wang, P. Luo. StyleAdapter: A unified stylized image generation model. *International Journal of Computer Vision*, vol. 133, no. 4, pp. 1894–1911, 2025. DOI: 10.1007/s11263-024-02253-x.

[30] X. Pan, L. Dong, S. Huang, Z. Peng, W. Chen, F. Wei. Kosmos-G: Generating images in context with multimodal large language models. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[31] S. Motamed, D. P. Paudel, L. Van Gool. LEGO: Learning to disentangle and invert personalized concepts beyond object appearance in text-to-image diffusion models. In *Proceedings of the 18th European Computer Vision Association*, Milan, Italy, pp. 116–133, 2025. DOI: 10.1007/978-3-031-72633-0_7.

[32] Y. Yan, C. Zhang, R. Wang, Y. Zhou, G. Zhang, P. Cheng, G. Yu, B. Fu. FaceStudio: Put your face everywhere in seconds, [Online], Available: https://arxiv.org/abs/2312.02663, 2023.

[33] Z. Li, M. Cao, X. Wang, Z. Qi, M. M. Cheng, Y. Shan. PhotoMaker: Customizing realistic human photos via stacked ID embedding. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8640–8650, 2024. DOI: 10.1109/CVPR52733.2024.00825.

[34] X. Peng, J. Zhu, B. Jiang, Y. Tai, D. Luo, J. Zhang, W. Lin, T. Jin, C. Wang, R. Ji. PortraitBooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 27070–27080, 2024. DOI: 10.1109/CVPR52733.2024.02557.

[35] X. Zhang, X. Y. Wei, J. Wu, T. Zhang, Z. Zhang, Z. Lei, Q. Li. Compositional inversion for stable diffusion models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 7350–7358, 2024. DOI: 10.1609/aaai.v38i7.28565.

[36] Q. Sun, Y. Cui, X. Zhang, F. Zhang, Q. Yu, Y. Wang, Y. Rao, J. Liu, T. Huang, X. Wang. Generative multimodal models are in-context learners. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 14398–14409, 2024. DOI: 10.1109/CVPR52733.2024.01365.

[37] L. Pang, J. Yin, H. Xie, Q. Wang, Q. Li, X. Mao. Cross initialization for personalized text-to-image generation, [Online], Available: https://arxiv.org/abs/2312.15905, 2023.

[38] Z. Kong, Y. Zhang, T. Yang, T. Wang, K. Zhang, B. Wu, G. Chen, W. Liu, W. Luo. OMG: Occlusion-

friendly personalized multi-concept generation in diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 253–270, 2025. DOI: 10.1007/978-3-031-72751-1_15.

[39] X. Zhang, W. Zhang, X. Wei, J. Wu, Z. Zhang, Z. Lei, Q. Li. Generative active learning for image synthesis personalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne, Australia, pp. 10669–10677, 2024. DOI: 10.1145/3664647.3680773.

[40] Y. Zhang, Y. Song, J. Liu, R. Wang, J. Yu, H. Tang, H. Li, X. Tang, Y. Hu, H. Pan, Z. Jiang. SSR-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8069–8078, 2024. DOI: 10.1109/CVPR52733.2024.00771.

[41] G. Zhang, K. Sohn, M. Hahn, H. Shi, I. Essa. FineStyle: Fine-grained controllable style personalization for text-to-image models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 52937–52961, 2024.

[42] Z. Dong, P. Wei, L. Lin. DreamArtist++: Controllable one-shot text-to-image generation via positive-negative adapter, [Online], Available: https://arxiv.org/abs/2211.11337, 2025.

[43] A. Voronov, M. Khoroshikh, A. Babenko, M. Ryabinin. Is this loss informative? Faster text-to-image customization by tracking objective dynamics. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1630, 2023.

[44] I. Han, S. Yang, T. Kwon, J. C. Ye. Highly personalized text embedding for image manipulation by stable diffusion, [Online], Available: https://arxiv.org/abs/2303.08767, 2023.

[45] C. Xiang, F. Bao, C. Li, H. Su, J. Zhu. A closer look at parameter-efficient tuning in diffusion models, [Online], Available: https://arxiv.org/abs/2303.18181, 2023.

[46] X. Jia, Y. Zhao, K. C. K. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, Y. C. Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models, [Online], Available: https://arxiv.org/abs/2304.02642, 2023.

[47] J. Yang, H. Wang, Y. Zhang, R. Xiao, S. Wu, G. Chen, J. Zhao. Controllable textual inversion for personalized text-to-image generation, [Online], Available: https://arxiv.org/abs/2304.05265, 2023.

[48] Z. Fei, M. Fan, J. Huang. Gradient-free textual inversion. In *Proceedings of the 31st ACM International Conference on Multimedia*, Ottawa, Canada, pp. 1364–1373, 2023. DOI: 10.1145/3581783.3612599.

[49] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, Y. Qiao, P. Gao, H. Li. Personalize segment anything model with one shot. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[50] O. Avrahami, K. Aberman, O. Fried, D. Cohen-Or, D. Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *Proceedings of SIGGRAPH Asia Conference Papers*, Sydney, Australia, Article number 96, 2023. DOI: 10.1145/3610548.3618154.

[51] J. Xiao, M. Yin, Y. Gong, X. Zang, J. Ren, B. Yuan. COMCAT: Towards efficient compression and customization of attention-based vision models. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, Article number 1587, 2023.

[52] S. Hao, K. Han, S. Zhao, K. Y. K. Wong. ViCo: Plug-and-play visual condition for personalized text-to-image generation, [Online], Available: https://arxiv.org/abs/2306.00971, 2023.

[53] Z. Qiu, W. Liu, H. Feng, Y. Xue, Y. Feng, Z. Liu, D. Zhang, A. Weller, B. Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 79320–79362, 2023.

[54] S. Y. Yeh, Y. G. Hsieh, Z. Gao, B. B. W. Yang, G. Oh, Y. Gong. Navigating text-to-image customization: From LyCORIS fine-tuning to model evaluation, [Online], Available: https://arxiv.org/abs/2309.14859, 2024.

[55] X. He, Z. Cao, N. Kolkin, L. Yu, K. Wan, H. Rhodin, R. Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Tucson, USA, pp. 3782–3791, 2025. DOI: 10.1109/WACV61041.2025.00372.

[56] A. Roy, M. Suin, A. Shah, K. Shah, J. Liu, R. Chellappa. DIFFNAT: Improving diffusion image quality using natural image statistics, [Online], Available: https://arxiv.org/abs/2311.09753, 2023.

[57] A. Agarwal, S. Karanam, T. Shukla, B. V. Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Tucson, USA, pp. 6053–6062, 2025. DOI: 10.1109/WACV61041.2025.00590.

[58] R. Zhao, M. Zhu, S. Dong, D. Cheng, N. Wang, X. Gao. CatVersion: Concatenating embeddings for diffusion-based text-to-image personalization. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 6, pp. 6047–6058, 2025. DOI: 10.1109/TCSVT.2025.3531917.

[59] M. Safaee, A. Mikaeili, O. Patashnik, D. Cohen-Or, A. Mahdavi-Amiri. CLiC: Concept learning in context. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6924–6933, 2024. DOI: 10.1109/CVPR52733.2024.00661.

[60] Z. Wang, W. Wei, Y. Zhao, Z. Xiao, M. Hasegawa-Johnson, H. Shi, T. Hou. HiFi tuner: High-fidelity subject-driven fine-tuning for diffusion models, [Online], Available: https://arxiv.org/abs/2312.00079, 2023.

[61] D. Chae, N. Park, J. Kim, K. Lee. InstructBooth: Instruction-following personalized text-to-image generation, [Online], Available: https://arxiv.org/abs/2312.03011, 2024.

[62] Y. Cai, Y. Wei, Z. Ji, J. Bai, H. Han, W. Zuo. Decoupled textual embeddings for customized image generation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 909–917, 2024. DOI: 10.1609/aaai.v38i2.27850.

[63] B. N. Zhao, Y. Xiao, J. Xu, X. Jiang, Y. Yang, D. Li, L. Itti, V. Vineet, Y. Ge. DreamDistribution: Learning prompt distribution for diverse in-distribution generation, [Online], Available: https://arxiv.org/abs/2312.14216, 2025.

[64] M. Hua, J. Liu, F. Ding, W. Liu, J. Wu, Q. He. DreamTuner: Single image is enough for subject-driven genera-

tion, [Online], Available: https://arxiv.org/abs/2312.13691, 2023.

[65] J. Lu, C. Xie, H. Guo. Object-driven one-shot fine-tuning of text-to-image diffusion with prototypical embedding, [Online], Available: https://arxiv.org/abs/2401.15708, 2024.

[66] W. Zeng, Y. Yan, Q. Zhu, Z. Chen, P. Chu, W. Zhao, X. Yang. Infusion: Preventing customized text-to-image diffusion from overfitting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne, Australia, pp. 3568–3577, 2024. DOI: 10.1145/3664647.3680894.

[67] M. Jones, S. Y. Wang, N. Kumari, D. Bau, J. Y. Zhu. Customizing text-to-image models with a single image pair, [Online], Available: https://arxiv.org/abs/2405.01536, 2024.

[68] H. Chen, Y. Zhang, X. Wang, X. Duan, Y. Zhou, W. Zhu. DisenDreamer: Subject-driven text-to-image generation with sample-aware disentangled tuning. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6860–6873, 2024. DOI: 10.1109/TCSVT.2024.3369757.

[69] Z. Fan, Z. Yin, G. Li, Y. Zhan, H. Zheng. DreamBooth++: Boosting subject-driven generation via region-level references packing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne, Australia, pp. 11013–11021, 2024. DOI: 10.1145/3664647.3680734.

[70] F. Wu, Y. Pang, J. Zhang, L. Pang, J. Yin, B. Zhao, Q. Li, X. Mao. CoRe: Context-regularized text embedding learning for text-to-image personalization, [Online], Available: https://arxiv.org/abs/2408.15914, 2024.

[71] J. Jin, Y. Shen, Z. Fu, J. Yang. Customized generation reimagined: Fidelity and editability harmonized. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 410–426, 2025. DOI: 10.1007/978-3-031-72973-7_24.

[72] W. Chen, H. Hu, C. Saharia, W. W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[73] X. Xu, Z. Wang, E. Zhang, K. Wang, H. Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 7720–7731, 2023. DOI: 10.1109/ICCV51070.2023.00713.

[74] R. Gal, M. Arar, Y. Atzmon, A. H. Bermano, G. Chechik, D. Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics*, vol. 42, no. 4, Article number 150, 2023. DOI: 10.1145/3592133.

[75] Y. Ma, H. Yang, W. Wang, J. Fu, J. Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation, [Online], Available: https://arxiv.org/abs/2303.09319, 2023.

[76] M. Kim, J. Yoo, S. Kwon. Personalized text-to-image model enhancement strategies: SOD preprocessing and CNN local feature integration. *Electronics*, vol. 12, no. 22, Article number 4707, 2023. DOI: 10.3390/electronics12224707.

[77] Y. Zhou, R. Zhang, J. Gu, T. Sun. Customization assistant for text-to-image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 9182–9191, 2024. DOI: 10.1109/CVPR52733.2024.00877.

[78] J. Pan, H. Yan, J. H. Liew, J. Feng, V. Y. F. Tan. Towards accurate guided diffusion sampling through symplectic adjoint method, [Online], Available: https://arxiv.org/abs/2312.12030, 2023.

[79] S. Purushwalkam, A. Gokul, S. Joty, N. Naik. BootPIG: Bootstrapping zero-shot personalized image generation capabilities in pretrained diffusion models. In *Proceedings of Computer Vision*, Milan, Italy, pp. 252–269, 2025. DOI: 10.1007/978-3-031-91907-7_15.

[80] N. Chen, M. Huang, Z. Chen, Y. Zheng, L. Zhang, Z. Mao. CustomContrast: A multilevel contrastive perspective for subject-driven text-to-image customization. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, USA, pp. 2123–2131, 2025. DOI: 10.1609/aaai.v39i2.32210.

[81] K. Song, Y. Zhu, B. Liu, Q. Yan, A. Elgammal, X. Yang. MoMA: Multimodal LLM adapter for fast personalized image generation. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 117–132, 2025. DOI: 10.1007/978-3-031-73661-2_7.

[82] C. Shin, J. Choi, H. Kim, S. Yoon. Large-scale text-to-image model with inpainting is a zero-shot subject-driven image generator, [Online], Available: https://arxiv.org/abs/2411.15466, 2024.

[83] J. Park, B. Ko, H. Jang. StyleBoost: A study of personalizing text-to-image generation in any style using dreambooth. In *Proceedings of the 14th International Conference on Information and Communication Technology Convergence*, Jeju Island, Republic of Korea, pp. 93–98, 2023. DOI: 10.1109/ICTC58733.2023.10392676.

[84] A. Hertz, A. Voynov, S. Fruchter, D. Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 4775–4785, 2024. DOI: 10.1109/CVPR52733.2024.00457.

[85] J. Park, B. Ko, H. Jang. Text-to-image synthesis for any artistic styles: Advancements in personalized artistic image generation via subdivision and dual binding, [Online], Available: https://arxiv.org/html/2404.05256v1, 2024.

[86] J. Choi, C. Shin, Y. Oh, H. Kim, J. Lee, S. Yoon. Style-friendly SNR sampler for style-driven generation, [Online], Available: https://arxiv.org/abs/2411.14793, 2024.

[87] D. Y. Chen, H. Tennent, C. W. Hsu. ArtAdapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*,Seattle,USA,pp. 8619–8628,2024.DOI:10.1109/CVPR52733.2024.00823.

[88] Y. Zhou, R. Zhang, T. Sun, J. Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach, [Online], Available: https://arxiv.org/abs/2305.13579, 2023.

[89] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, N. Memon. Identity-preserving aging of face images via latent diffusion models. In *Proceedings of IEEE International Joint Conference on Biometrics*, Ljubljana, Slovenia, 2023. DOI: 10.1109/IJCB57857.2023.10448860.

[90] J. Hyung, J. Shin, J. Choo. MagiCapture: High-resolution multi-concept portrait customization, [Online], Available: https://arxiv.org/abs/2309.06895, 2024.

[91] P. Cao, L. Yang, F. Zhou, T. Huang, Q. Song. Concept-centric personalization with large-scale diffusion priors, [Online], Available: https://arxiv.org/html/2312.08195v1, 2023.

[92] C. Kim, J. Lee, S. Joung, B. Kim, Y. M. Baek. InstantFamily: Masked attention for zero-shot multi-ID image generation, [Online], Available: https://arxiv.org/abs/2404.19427, 2024.

[93] X. Li, J. Zhan, S. He, Y. Xu, J. Dong, H. Zhang, Y. Du. PersonaMagic: Stage-regulated high-fidelity face customization with tandem equilibrium, [Online], Available: https://arxiv.org/abs/2412.15674, 2024.

[94] Y. C. Su, K. C. K. Chan, Y. Li, Y. Zhao, H. Zhang, B. Gong, H. Wang, X. Jia. Identity encoder for personalized diffusion, [Online], Available: https://arxiv.org/abs/2304.07429, 2023.

[95] Z. Chen, S. Fang, W. Liu, Q. He, M. Huang, Y. Zhang, Z. Mao. DreamIdentity: Improved editability for efficient face-identity preserved image generation, [Online], Available: https://arxiv.org/abs/2307.00300, 2023.

[96] Y. Wang, W. Zhang, J. Zheng, C. Jin. High-fidelity person-centric subject-to-image synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 7675–7684, 2024. DOI: 10.1109/CVPR52733.2024.00733.

[97] X. Li, X. Hou, C. C. Loy. When StyleGAN meets stable diffusion: A W_+ adapter for personalized image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 2187–2196, 2024. DOI: 10.1109/CVPR52733.2024.00213.

[98] J. Liu, H. Huang, C. Jin, R. He. Portrait diffusion: Training-free face stylization with chain-of-painting, [Online], Available: https://arxiv.org/abs/2312.02212, 2023.

[99] H. Tang, J. Deng, Z. Pan, H. Tian, P. Chaudhari, X. Zhou. RetriBooru: Leakage-free retrieval of conditions from reference images for subject-driven generation, [Online], Available: https://arxiv.org/abs/2312.02521, 2024.

[100] J. Xu, S. Motamed, P. Vaddamanu, C. H. Wu, C. Haene, J. C. Bazin, F. De La Torre. Personalized face inpainting with diffusion models by parallel visual attention. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 5420–5430, 2024. DOI: 10.1109/WACV57701.2024.00535.

[101] D. Y. Chen, A. K. Bhunia, S. Koley, A. Sain, P. N. Chowdhury, Y. Z. Song. DemoCaricature: Democratising caricature generation with a rough sketch. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8629–8639, 2024. DOI: 10.1109/CVPR52733.2024.00824.

[102] P. Achlioptas, A. Benetatos, I. Fostiropoulos, D. Skourtis. Stellar: Systematic evaluation of human-centric personalized text-to-image methods, [Online], Available: https://arxiv.org/abs/2312.06116, 2023.

[103] W. Chen, J. Zhang, J. Wu, H. Wu, X. Xiao, L. Lin. ID-aligner: Enhancing identity-preserving text-to-image generation with reward feedback learning, [Online], Available: https://arxiv.org/abs/2404.15449, 2024.

[104] K. C. Wang, D. Ostashev, Y. Fang, S. Tulyakov, K. Aberman. MoA: Mixture-of-attention for subject-context disentanglement in personalized image generation, [Online], Available: https://arxiv.org/html/2404.11565v1, 2024.

[105] S. Cui, J. Guo, X. An, J. Deng, Y. Zhao, X. Wei, Z. Feng. IDAdapter: Learning mixed features for tuning-free personalization of text-to-image models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, USA, pp. 950–959, 2024. DOI: 10.1109/CVPRW63382.2024.00100.

[106] Y. Wu, Z. Li, H. Zheng, C. Wang, B. Li. Infinite-ID: Identity-preserved personalization via ID-semantics decoupling paradigm. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 279–296, 2025. DOI: 10.1007/978-3-031-73242-3_16.

[107] K. Shiohara, T. Yamasaki. Face2Diffusion for fast and editable face personalization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6850–6859, 2024. DOI: 10.1109/CVPR52733.2024.00654.

[108] Y. Cai, Z. Jiang, Y. Liu, C. Jiang, W. Xue, W. Luo, Y. Guo. Foundation cures personalization: Recovering facial personalized models' prompt consistency, [Online], Available: https://arxiv.org/abs/2411.15277v1, 2024.

[109] G. Qian, K. C. Wang, O. Patashnik, N. Heravi, D. Ostashev, S. Tulyakov, D. Cohen-Or, K. Aberman. Omni-ID: Holistic identity representation designed for generative tasks, [Online], Available: https://arxiv.org/html/2412.09694v1, 2024.

[110] Y. Li, H. Liu, Y. Wen, Y. J. Lee. Generate anything anywhere in any scene, [Online], Available: https://arxiv.org/abs/2306.17154, 2023.

[111] T. Rahman, S. Mahajan, H. Y. Lee, J. Ren, S. Tulyakov, L. Sigal. Visual concept-driven image generation with text-to-image diffusion model, [Online], Available: https://arxiv.org/abs/2402.11487, 2025.

[112] J. Jiang, Y. Zhang, K. Feng, X. Wu, W. Li, R. Pei, F. Li, W. Zuo. MC^2: Multi-concept guidance for customized multi-concept generation, [Online], Available: https://arxiv.org/abs/2404.05268, 2024.

[113] C. Zhu, K. Li, Y. Ma, C. He, X. Li. MultiBooth: Towards generating all your concepts in an image from text. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, USA, pp. 10923–10931, 2025. DOI: 10.1609/aaai.v39i10.33187.

[114] H. Matsuda, R. Togo, K. Maeda, T. Ogawa, M. Haseyama. Multi-object editing in personalized text-to-image diffusion model via segmentation guidance. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, Republic of Korea, pp. 8140–8144, 2024. DOI: 10.1109/ICASSP48485.2024.10447048.

[115] D. Zhou, J. Huang, J. Bai, J. Wang, H. Chen, G. Chen, X. Hu, P. A. Heng. MagicTailor: Component-controllable personalization in text-to-image diffusion models, [Online], Available: https://arxiv.org/abs/2410.13370, 2024.

[116] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, H. Zhao. AnyDoor: Zero-shot object-level image customization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6593–6602, 2024. DOI: 10.1109/CVPR52733.2024.00630.

[117] Z. Yuan, M. Cao, X. Wang, Z. Qi, C. Yuan, Y. Shan. CustomNet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models, [Online], Available: https://arxiv.org/abs/2310.19784, 2023.

[118] D. Zhou, Y. Li, F. Ma, Z. Yang, Y. Yang. MIGC: Multi-instance generation controller for text-to-image

synthesis. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6818–6828, 2024. DOI: 10.1109/CVPR52733.2024.00651.

[119] M. Patel, S. Jung, C. Baral, Y. Yang. λ-ECLIPSE: Multi-concept personalized text-to-image diffusion models by leveraging CLIP latent space, [Online], Available: https://arxiv.org/abs/2402.05195, 2024.

[120] X. Wang, S. Fu, Q. Huang, W. He, H. Jiang. MS-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, 2025.

[121] Z. Huang, T. Wu, Y. Jiang, K. C. K. Chan, Z. Liu. ReVersion: Diffusion-based relation inversion from images. In *Proceedings of SIGGRAPH Asia Conference Papers*, Tokyo, Japan, Article number 4, 2024. DOI: 10.1145/3680528.3687658.

[122] G. Zhang, Y. Qian, J. Deng, X. Cai. Inv-ReVersion: Enhanced relation inversion based on text-to-image diffusion models. *Applied Sciences*, vol. 14, no. 8, Article number 3338, 2024. DOI: 10.3390/app14083338.

[123] Z. Xu, S. Hao, K. Han. CusConcept: Customized visual concept decomposition with diffusion models. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Tucson, USA, pp. 3678–3687, 2025. DOI: 10.1109/WACV61041.2025.00362.

[124] S. Huang, B. Gong, Y. Feng, X. Chen, Y. Fu, Y. Liu, D. Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 7797–7806, 2024. DOI: 10.1109/CVPR52733.2024.00745.

[125] J. Gu, Y. Wang, N. Zhao, T. J. Fu, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung, X. E. Wang. PHOTOSWAP: Personalized subject swapping in images. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1529, 2023.

[126] S. Zhang, M. Ni, S. Chen, L. Wang, W. Ding, Y. Liu. A two-stage personalized virtual try-on framework with shape control and texture guidance. *IEEE Transactions on Multimedia*, vol. 26, pp. 10225–10236, 2024. DOI: 10.1109/TMM.2024.3405718.

[127] M. Chen, I. Laina, A. Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 5331–5341, 2024. DOI: 10.1109/WACV57701.2024.00526.

[128] J. Gu, N. Zhao, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung, Y. Wang, X. E. Wang. SwapAnything: Enabling arbitrary object swapping in personalized visual editing, [Online], Available: https://arxiv.org/abs/2404.05717, 2024.

[129] N. Kumari, G. Su, R. Zhang, T. Park, E. Shechtman, J. Y. Zhu. Customizing text-to-image diffusion with object viewpoint control, [Online], Available: https://arxiv.org/abs/2404.12333, 2024.

[130] X. Xu, J. Guo, Z. Wang, G. Huang, I. Essa, H. Shi. Prompt-free diffusion: Taking "text" out of text-to-image diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 8682–8692, 2024. DOI: 10.1109/CVPR52733.2024.00829.

[131] S. Zhao, D. Chen, Y. C. Chen, J. Bao, S. Hao, L. Yuan, K. Y. K. Wong. Uni-controlNet: All-in-one control to text-to-image diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 11127–11150, 2023.

[132] S. Y. Cheong, A. Mustafa, A. Gilbert. ViscoNet: Bridging and harmonizing visual and textual conditioning for ControlNet, [Online], Available: https://arxiv.org/abs/2312.03154, 2024.

[133] I. Najdenkoska, A. Sinha, A. Dubey, D. Mahajan, V. Ramanathan, F. Radenovic. Context diffusion: In-context aware image generation. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 375–391, 2024. DOI: 10.1007/978-3-031-72980-5_22.

[134] S. Mo, F. Mu, K. H. Lin, Y. Liu, B. Guan, Y. Li, B. Zhou. FreeControl: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 7465–7475, 2024. DOI: 10.1109/CVPR52733.2024.00713.

[135] P. Li, Q. Nie, Y. Chen, X. Jiang, K. Wu, Y. Lin, Y. Liu, J. Peng, C. Wang, F. Zheng. Tuning-free image customization with image and text guidance. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 233–250, 2025. DOI: 10.1007/978-3-031-73116-7_14.

[136] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, M. Z. Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 7623–7633, 2023. DOI: 10.1109/ICCV51070.2023.00701.

[137] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, A. Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 7346–7356, 2023. DOI: 10.1109/ICCV51070.2023.00675.

[138] Y. Zhao, E. Xie, L. Hong, Z. Li, G. H. Lee. Make-a-protagonist: Generic video editing with an ensemble of experts, [Online], Available: https://ar5iv.labs.arxiv.org/html/2305.08850, 2023.

[139] Y. He, M. Xia, H. Chen, X. Cun, Y. Gong, J. Xing, Y. Zhang, X. Wang, C. Weng, Y. Shan, Q. Chen. Animate-a-story: Storytelling with retrieval-augmented video generation, [Online], Available: https://arxiv.org/abs/2307.06940, 2023.

[140] R. Zhao, Y. Gu, J. Z. Wu, D. J. Zhang, J. W. Liu, W. Wu, J. Keppo, M. Z. Shou. MotionDirector: Motion customization of text-to-video diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 273–290, 2025. DOI: 10.1007/978-3-031-72992-8_16.

[141] R. Wu, L. Chen, T. Yang, C. Guo, C. Li, X. Zhang. LAMP: Learn a motion pattern for few-shot video generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 7089–7098, 2024. DOI: 10.1109/CVPR52733.2024.00677.

[142] H. Jeong, G. Y. Park, J. C. Ye. VMC: Video motion customization using temporal attention adaption for

text-to-video diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 9212–9221, 2024. DOI: 10.1109/CVPR52733.2024.00880.

[143] Y. Song, W. Shin, J. Lee, J. Kim, N. Kwak. SAVE: Protagonist diversification with structure agnostic video editing. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 41–57, 2025. DOI: 10.1007/978-3-031-72989-8_3.

[144] J. Materzynska, J. Sivic, E. Shechtman, A. Torralba, R. Zhang, B. Russell. NewMove: Customizing text-to-video models with novel motions, [Online], Available: https://arxiv.org/abs/2312.04966, 2023.

[145] Y. Wei, S. Zhang, Z. Qing, H. Yuan, Z. Liu, Y. Liu, Y. Zhang, J. Zhou, H. Shan. DreamVideo: Composing your dream videos with customized subject and motion, [Online], Available: https://arxiv.org/abs/2312.04433, 2023.

[146] Y. Zhang, F. Tang, N. Huang, H. Huang, C. Ma, W. Dong, C. Xu. MotionCrafter: One-shot motion customization of diffusion models, [Online], Available: https://arxiv.org/abs/2312.05288, 2024.

[147] Y. Ren, Y. Zhou, J. Yang, J. Shi, D. Liu, F. Liu, M. Kwon, A. Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 332–349, 2025. DOI: 10.1007/978-3-031-73024-5_20.

[148] Z. Ma, D. Zhou, C. H. Yeh, X. S. Wang, X. Li, H. Yang, Z. Dong, K. Keutzer, J. Feng. Magic-Me: Identity-specific video customized diffusion, [Online], Available: https://arxiv.org/abs/2402.09368, 2024.

[149] X. Bi, J. Lu, B. Liu, X. Cun, Y. Zhang, W. Li, B. Xiao. CustomTTT: Motion and appearance customized video generation via test-time training. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, USA, pp. 1871–1879, 2025. DOI: 10.1609/aaai.v39i2.32182.

[150] H. Li, H. Qiu, S. Zhang, X. Wang, Y. Wei, Z. Li, Y. Zhang, B. Wu, D. Cai. PersonalVideo: High ID-fidelity video customization without dynamic and semantic degradation, [Online], Available: https://arxiv.org/abs/2411.17048, 2024.

[151] Z. Wang, A. Li, L. Zhu, Y. Guo, Q. Dou, Z. Li. CustomVideo: Customizing text-to-video generation with multiple subjects, [Online], Available: https://arxiv.org/abs/2401.09962, 2024.

[152] H. Chen, X. Wang, G. Zeng, Y. Zhang, Y. Zhou, F. Han, Y. Wu, W. Zhu. VideoDreamer: Customized multi-subject text-to-video generation with disen-mix finetuning on language-video foundation models, [Online], Available: https://arxiv.org/abs/2311.00990, 2023.

[153] H. Zhao, T. Lu, J. Gu, X. Zhang, Q. Zheng, Z. Wu, H. Xu, Y. G. Jiang. MagDiff: Multi-alignment diffusion for high-fidelity video generation and editing. In *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 205–221, 2025. DOI: 10.1007/978-3-031-72649-1_12.

[154] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, Z. Liu. VideoBooth: Diffusion-based video generation with image prompts. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 6689–6700, 2024. DOI: 10.1109/CVPR52733.2024.00639.

[155] T. Wu, Y. Zhang, X. Cun, Z. Qi, J. Pu, H. Dou, G. Zheng, Y. Shan, X. Li. VideoMaker: Zero-shot customized video generation with the inherent force of video diffusion models, [Online], Available: https://arxiv.org/abs/2412.19645, 2024.

[156] Y. Zhou, R. Zhang, J. Gu, N. Zhao, J. Shi, T. Sun. SUGAR: Subject-driven video customization in a zero-shot manner, [Online], Available: https://arxiv.org/abs/2412.10533, 2024.

[157] G. Liu, M. Xia, Y. Zhang, H. Chen, J. Xing, Y. Wang, X. Wang, Y. Shan, Y. Yang. StyleCrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Transactions on Graphics*, vol. 43, no. 6, Article number 251, 2024. DOI: 10.1145/3687975

[158] X. He, Q. Liu, S. Qian, X. Wang, T. Hu, K. Cao, K. Yan, J. Zhang. ID-animator: Zero-shot identity-preserving human video generation, [Online], Available: https://arxiv.org/abs/2404.15275, 2024.

[159] S. Yuan, J. Huang, X. He, Y. Ge, Y. Shi, L. Chen, J. Luo, L. Yuan. Identity-preserving text-to-video generation by frequency decomposition, [Online], Available: https://arxiv.org/abs/2411.17440, 2024.

[160] H. Fang, D. Qiu, B. Mao, P. Yan, H. Tang. MotionCharacter: Identity-preserving and motion controllable human video generation, [Online], Available: https://arxiv.org/abs/2411.18281, 2024.

[161] M. Feng, J. Liu, K. Yu, Y. Yao, Z. Hui, X. Guo, X. Lin, H. Xue, C. Shi, X. Li, A. Li, X. Kang, B. Lei, M. Cui, P. Ren, X. Xie. DreaMoving: A human video generation framework based on diffusion models, [Online], Available: https://arxiv.org/abs/2312.05107, 2023.

[162] C. H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M. Y. Liu, T. Y. Lin. Magic3D: High-resolution text-to-3D content creation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 300–309, 2023. DOI: 10.1109/CVPR52729.2023.00037.

[163] A. Raj, S. Kaza, B. Poole, M. Niemeyer, N. Ruiz, B. Mildenhall, S. Zada, K. Aberman, M. Rubinstein, J. Barron, Y. Li, V. Jampani. DreamBooth3D: Subject-driven text-to-3D generation. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 2349–2359, 2023. DOI: 10.1109/ICCV51070.2023.00223.

[164] S. Azadi, T. Hayes, A. Shah, G. Pang, D. Parikh, S. Gupta. Text-conditional contextualized avatars for zero-shot personalization, [Online], Available: https://arxiv.org/abs/2304.07410, 2023.

[165] C. Zhang, Y. Chen, Y. Fu, Z. Zhou, G. Yu, B. Wang, B. Fu, T. Chen, G. Lin, C. Shen. StyleAvatar3D: Leveraging image-text diffusion models for high-fidelity 3D avatar generation, [Online], Available: https://arxiv.org/abs/2305.19012, 2023.

[166] Y. Zeng, Y. Lu, X. Ji, Y. Yao, H. Zhu, X. Cao. AvatarBooth: High-quality and customizable 3D human avatar generation, [Online], Available: https://arxiv.org/abs/2306.09864, 2023.

[167] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, X. Yang. MVDream: Multi-view diffusion for 3D generation. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[168] Y. Ouyang, W. Chai, J. Ye, D. Tao, Y. Zhan, G. Wang.

Chasing consistency in text-to-3D generation from a single image, [Online], Available: https://arxiv.org/abs/2309.03599, 2023.

[169] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, G. H. Lee. Animate124: Animating one image to 4D dynamic scene, [Online], Available: https://arxiv.org/abs/2311.14603, 2023.

[170] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, S. De Mello. A unified approach for text-and image-guided 4D scene generation. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 7300–7309, 2024. DOI: 10.1109/CVPR52733.2024.00697.

[171] Y. Y. Yeh, J. B. Huang, C. Kim, L. Xiao, T. Nguyen-Phuoc, N. Khan, C. Zhang, M. Chandraker, C. S. Marshall, Z. Dong, Z. Li. TextureDreamer: Image-guided texture synthesis through geometry-aware diffusion. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 4304–4314, 2024. DOI: 10.1109/CVPR52733.2024.00412.

[172] J. Zhuang, D. Kang, Y. P. Cao, G. Li, L. Lin, Y. Shan. TIP-editor: An accurate 3D editor following both text-prompts and image-prompts. *ACM Transactions on Graphics*, vol. 43, no. 4, Article number 121, 2024. DOI: 10.1145/3658205.

[173] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, A. Tran. Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 2116–2127, 2023. DOI: 10.1109/ICCV51070.2023.00202.

[174] Y. Wu, J. Zhang, F. Kerschbaum, T. Zhang. Backdooring textual inversion for concept censorship, [Online], Available: https://arxiv.org/abs/2308.10718, 2023.

[175] Y. Huang, F. Juefei-Xu, Q. Guo, J. Zhang, Y. Wu, M. Hu, T. Li, G. Pu, Y. Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, pp. 21169–21178, 2024. DOI: 10.1609/aaai.v38i19.30110.

[176] J. S. Smith, Y. C. Hsu, L. Zhang, T. Hua, Z. Kira, Y. Shen, H. Jin. Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA, [Online], Available: https://arxiv.org/abs/2304.06027, 2024.

[177] P. Zhang, N. Zhao, J. Liao. Text-guided vector graphics customization. In *Proceedings of SIGGRAPH Asia Conference Papers*, Sydney, Australia, Article number 54, 2023. DOI: 10.1145/3610548.3618232.

[178] H. Wang, X. Xiang, Y. Fan, J. H. Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, USA, pp. 4921–4931, 2024. DOI: 10.1109/WACV57701.2024.00486.

[179] K. Wang, F. Yang, B. Raducanu, J. van de Weijer. Multi-class textual-inversion secretly yields a semantic-agnostic classifier. In *Proceedings of IEEE/CVF Winter Conference on Applications of Computer Vision*, Tucson, USA, pp. 4400–4409, 2025. DOI: 10.1109/WACV61041.2025.00432.

[180] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[181] T. Karras, M. Aittala, T. Aila, S. Laine. Elucidating the design space of diffusion-based generative models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 26565–26577, 2022.

[182] J. Ho, A. Jain, P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 6840–6851, 2020.

[183] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982. DOI: 10.1016/0304-4149(82)90051-5.

[184] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, pp. 5775–5787, 2022.

[185] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu. DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Reserach*, vol. 22, no. 4, pp. 730–751, 2025. DOI: 10.1007/s11633-025-1562-4.

[186] D. Chen, Z. Zhou, C. Wang, C. Shen, S. Lyu. On the trajectory regularity of ODE-based diffusion sampling. In *Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria, pp. 7905–7934, 2024.

[187] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 10684–10695, 2022. DOI: 10.1109/CVPR52688.2022.01042.

[188] L. Zhang, A. Rao, M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 3813–3824, 2023. DOI: 10.1109/ICCV51070.2023.00355.

[189] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen. Hierarchical text-conditional image generation with CLIP latents, [Online], Available: https://arxiv.org/abs/2204.06125, 2022.

[190] O. Ronneberger, P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[191] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, pp. 6000–6010, 2017.

[192] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, A. Ramesh. Improving image generation with better captions, [Online], Available: https://cdn.openai.com/papers/dall-e-3.pdf.

[193] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever. Learning transferable

visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

[194] J. Li, D. Li, C. Xiong, S. C. H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, Baltimore, USA, pp. 12888–12900, 2022.

[195] J. Li, D. Li, S. Savarese, S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, USA, Article number 814, 2023.

[196] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, J. Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1833, 2022.

[197] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, pp. 18676–18688, 2022. DOI: 10.1109/CVPR52688.2022.01814.

[198] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, A. Mian. Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, vol. 108, Article number 102417, 2024. DOI: 10.1016/j.inffus.2024.102417.

[199] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Y. Lo, P. Dollár, R. Girshick. Segment anything. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 3992–4003, 2023. DOI: 10.1109/ICCV51070.2023.00371.

[200] X. Y. Wei, Z. Q. Yang. Coaching the exploration and exploitation in active learning for interactive video retrieval. *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 955–968, 2013. DOI: 10.1109/TIP.2012.2222902.

[201] X. Y. Wei, Z. Q. Yang. Coached active learning for interactive video search. In *Proceedings of the 19th ACM International Conference on Multimedia*, Scottsdale, USA, pp. 443–452, 2011. DOI: 10.1145/2072298.2072356.

[202] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, X. Cao. FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 598–607, 2020. DOI: 10.1109/CVPR42600.2020.00068.

[203] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, Xi'an, China, pp. 67–74, 2018. DOI: 10.1109/FG.2018.00020.

[204] Y. Wu, Q. Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019. DOI: 10.1007/s11263-018-1097-z.

[205] I. Adjabi, A. Ouahabi, A. Benzaoui, A. Taleb-Ahmed. Past, present, and future of face recognition: A review. *Electronics*, vol. 9, no. 8, Article number 1188, 2020. DOI: 10.3390/electronics9081188.

[206] T. Karras, S. Laine, T. Aila. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021. DOI: 10.1109/tpami.2020.2970919.

[207] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

[208] J. Song, C. Meng, S. Ermon. Denoising diffusion implicit models. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

[209] T. Islam, A. Miron, X. Liu, Y. Li. Deep learning in virtual try-on: A comprehensive survey. *IEEE Access*, vol. 12, pp. 29475–29502, 2024. DOI: 10.1109/ACCESS.2024.3368612.

[210] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, Y. G. Jiang. A survey on video diffusion models. *ACM Computing Surveys*, vol. 57, no. 2, Article number 41, 2025. DOI: 10.1145/3696415.

[211] B. Poole, A. Jain, J. T. Barron, B. Mildenhall. DreamFusion: Text-to-3D using 2D diffusion, [Online], Available: https://arxiv.org/abs/2209.14988, 2022.

[212] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2022. DOI: 10.1145/3503250.

[213] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, Y. Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 700, 2023.

[214] X. Wu, K. Sun, F. Zhu, R. Zhao, H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 2096–2105, 2023. DOI: 10.1109/ICCV51070.2023.00200.

[215] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, H. Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis, [Online], Available: https://arxiv.org/abs/2306.09341, 2023.

[216] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, O. Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, USA, Article number 1594, 2023.

[217] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Montreal, Canada, pp. 9630–9640, 2021. DOI: 10.1109/ICCV48922.2021.00951.

[218] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceed-*

ings of the 31st International Conference on Neural Information Processing Systems, Long Beach, USA, pp. 6629–6640, 2017.

[219] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 2818–2826, 2016. DOI: 10.1109/CVPR.2016.308.

[220] Z. Liu, P. Luo, X. Wang, X. Tang. Deep learning face attributes in the wild. In *Proceedings of IEEE/CVF International Conference on Computer Vision*, Santiago, Chile, pp. 3730–3738, 2015. DOI: 10.1109/ICCV.2015.425.

[221] K. Zhang, Z. Zhang, Z. Li, Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.

[222] F. Schroff, D. Kalenichenko, J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 815–823, 2015. DOI: 10.1109/CVPR.2015.7298682.

[223] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu, Y. Zhao, Y. Ao, X. Min, T. Li, B. Wu, B. Zhao, B. Zhang, L. Wang, G. Liu, Z. He, X. Yang, J. Liu, Y. Lin, T. Huang, Z. Wang. Emu3: Next-token prediction is all you need, [Online], Available: https://arxiv.org/abs/2409.18869, 2024.

[224] Gemini Team Google. Gemini: A family of highly capable multimodal models, [Online], Available: https://arxiv.org/abs/2312.11805, 2023.

Xulu Zhang received the B. Eng. and the M. Sc. degrees in computer science from Sichuan University, China in 2019 and 2022. He is a Ph. D. degree candidate from the Department of Computing, The Hong Kong Polytechnic University, China.

His research interests include image generation and active learning.

E-mail: compxulu.zhang@connect.polyu.hk
ORCID iD: 0000-0003-2473-460X

Xiaoyong Wei received the Ph. D. degree in computer science from the City University of Hong Kong, China in 2009, and has worked as a postdoctoral fellow in the University of California, USA from December 2013 to December 2015. He has been a professor and the head of Department of Computer Science, Sichuan University, China since 2010. He is an adjunct professor of Peng Cheng Laboratory, China, and a visiting professor of Department of Computing, The Hong Kong Polytechnic University, China. He is a senior member of IEEE, and has served as an Associate Editor of *Interdisciplinary Sciences: Computational Life Sciences* since 2020, the program Chair of ICMR 2019, ICIMCS 2012, and the technical committee member of over 20 conferences such as ICCV, CVPR, SIGKDD, ACM MM, ICME, and ICIP.

His research interests include multimedia computing, health computing, machine learning and large-scale data mining.

E-mail: x1wei@polyu.edu.hk (Corresponding author)
ORCID iD: 0000-0002-5706-5177

Wentao Hu received the B. Eng. degree in computer science from Shandong University, China in 2021, and the M. Sc. degree in computer science from Sun Yat-sen University, China in 2024. He is currently a Ph. D. degree candidate from the Department of Computing, The Hong Kong Polytechnic University, China.

His research interests include image generation and 3D reconstruction.

E-mail: wayne-wt.hu@connect.polyu.hk
ORCID iD: 0000-0002-2071-9341

Jinlin Wu received the B. Sc. degree in computer science from the University of Electronic Science and Technology of China, China in 2017, and the Ph. D. degree in computer science from the University of Chinese Academy of Sciences, China in 2022. He is an assistant research fellow at the Institute of Automation, Chinese Academy of Sciences (CAS), China. He has served as the principal investigator of a National Natural Science Foundation of China (NSFC) Youth Science Fund project and has participated in several other NSFC-funded projects. He has accumulated a solid research foundation in areas such as video analysis for security and medical video understanding. He has published over 30 high-quality academic papers, with more than 500 citations.

His research interests include object detection, image recognition, and video understanding.

E-mail: jinlin.wu@cair-cas.org.hk
ORCID iD: 0000-0001-7877-5728

Jiaxin Wu received the Ph. D. degree in computer science from the City University of Hong Kong, China in 2024. She is a postdoctoral fellow in the Department of Computing at The Hong Kong Polytechnic University, China.

Her research interests include multimedia retrieval, AI for Science (AI4Science), and natural language processing (NLP).

E-mail: jiaxwu@polyu.edu.hk
ORCID iD: 0000-0003-4074-3442

Wengyu Zhang is an undergraduate student in the Department of Computing, The Hong Kong Polytechnic University, China.

His research interests include natural language processing (NLP), AI for Science (AI4Science), and graph learning.

E-mail: wengyu.zhang@connect.polyu.hk
ORCID iD: 0009-0001-2347-4183

Zhaoxiang Zhang received the B. Sc. degree in computer science in Department of Electronic Science and Technology, University of Science and Technology of China, China in 2004, and the Ph. D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, China in 2009, respectively. From 2009 to 2015, he worked as a lecturer, associate professor, and later the deputy director of Department of Computer Application Technology at the Beihang University, China. Since July 2015, he has joined the Institute of Automation, Chinese Academy of Sciences, where he is currently a professor. Recently, he specifically focuses on deep learning models, biologically-inspired visual computing and human-like learning, and their applications on human analysis and scene understanding. He has published more than 200 papers in international journals and conferences, including reputable international journals such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *International Journal of Computer Vision*, *Journal of Machine Learning Research*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, and top-level international conferences like CVPR, ICCV, NIPS, ECCV, ICLR, AAAI, IJCAI and ACM MM. He has won the best paper awards in several conferences and championships in international competitions and his research has won the "Technical Innovation Award of the Chinese Association of Artificial Intelligence". He has served as the PC Chair or Area Chair of many international conferences like CVPR, ICCV, AAAI, IJCAI, ACM MM, ICPR and BICS. He is serving or has served as Associate Editor of reputable international journals like *International Journal of Computer Vision*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *Pattern Recognition* and *Neurocomputing*.

His research interests include pattern recognition, computer vision and machine learning.

E-mail: zhaoxiang.zhang@ia.ac.cn
ORCID iD: 0000-0003-2648-3875

Zhen Lei received the B. Sc. degree in automation from the University of Science and Technology of China, China in 2005, and the Ph. D. degree in computer vision from the Institute of Automation, Chinese Academy of Sciences, China in 2010, where he is currently a professor. He is IEEE Fellow, IAPR Fellow and AAIA Fellow. He has published over 200 papers in international journals and conferences with 33 000+ citations in Google Scholar and h-index 86. He was the program co-chair of IJCB2023, was competition co-chair of IJCB2022 and has served as Area Chairs for several conferences and is an Associate Editor for *IEEE Transactions on Information Forensics and Security*, *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *Pattern Recognition*, *Neurocomputing* and *IET Computer Vision* journals. He is the winner of 2019 IAPR Young Biometrics Investigator Award.

His research interests include computer vision, pattern recognition, image processing, and face recognition in particular.

E-mail: zhen.lei@ia.ac.cn (Corresponding author)
ORCID iD: 0000-0002-0791-189X

Qing Li received the Ph. D. degree in data science from University of Southern California, USA in 1989. He is currently a Chair professor (Data Science) and the Head of the Department of Computing, The Hong Kong Polytechnic University. Formerly, he was the founding director of the Multimedia software Engineering Research Centre (MERC), and a professor at City University of Hong Kong where he worked in the Department of Computer Science from 1998 to 2018. Prior to these, he has also taught at the Hong Kong University of Science and Technology and the Australian National University (Canberra, Australia). He served as a consultant to Microsoft Research Asia (Beijing, China), Motorola Global Computing and Telecommunications Division (Tianjin Regional Operations Center), and the Division of Information Technology, Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia. He has been an adjunct professor of the University of Science and Technology of China (USTC) and the Wuhan University, and a guest professor of the Hunan University (Changsha, China) where he got his B. Eng. degree from the Department of Computer Science in 1982. He is also a guest professor (Software Technology) of the Zhejiang University (Hangzhou, China) – the leading university of the Zhejiang province where he was born.

He has been actively involved in the research community by serving as an Associate Editor and reviewer for technical journals, and as an organizer/co-organizer of numerous international conferences. Some recent conferences in which he is playing or has played major roles include APWeb-WAIM2018, ICDM 2018, WISE2017, ICDSC2016, DASFAA2015, U-Media2014, ER2013, RecSys2013, NDBC2012, ICMR2012, CoopIS2011, WAIM2010, DASFAA2010, APWeb-WAIM2009, ER2008, WISE2007, ICWL2006, HSI2005, WAIM2004, IDEAS2003, VLDB2002, PAKDD2001, IFIP 2.6 Working Conference on Database Semantics (DS-9), IDS2000, and WISE2000. In addition, he served as a programme committee member for over fifty international conferences (including VLDB, ICDE, WWW, DASFAA, ER, CIKM, CAiSE, CoopIS, and FODO). He is currently a fellow of IEEE and IET/IEE, a member of ACM-SIGMOD and IEEE Technical Committee on Data Engineering. He is the Chairperson of the Hong Kong Web Society, and also served/is serving as an executive committee (EXCO) member of IEEE-Hong Kong Computer Chapter and ACM Hong Kong Chapter. In addition, he serves as a councilor of the Database Society of Chinese Computer Federation (CCF), a member of the Big Data Expert Committee of CCF, and is a Steering Committee member of DASFAA, ER, ICWL, UMEDIA, and WISE Society.

His research interests include data science and artificial intelligence.

E-mail: qing-prof.li@polyu.edu.hk
ORCID iD: 0000-0003-3370-471X