Computational Visual Media
https://doi.org/10.26599/CVM.2025.9450511

Short Communication

Photorealistic fire scene video generation via multimodal large language model and pre-trained video diffusion model

Hongtao Zheng[1,2], **Xinyan Huang**[1] (✉)

Abstract Text-to-video diffusion models have made significant progress. However, there is still a lack of dedicated research on generating fire scene videos with physical realism and visual fidelity. To address this gap, we propose text-to-video fire (T2VFire) scene generation. T2VFire uses GPT-4o as the core engine, which is integrated with an external fire-related knowledge base and a retrieval-augmented generation (RAG) mechanism that can be dynamically updated based on prompts. With the support of this knowledge, the system first expands the user's initial text description and generates a keyframe image. Then, through iterative prompt optimization, it guides a pretrained video diffusion model to generate fire scene videos with physical consistency. Experimental results show that T2VFire improves upon the physical consistency and visual realism of fire scene videos generated by current video generation models. This method provides a solid foundation for future smart firefighting and digital twin systems in building fire safety management.

Keywords text-to-video (T2V); diffusion models; fire; video; physicality

1 Introduction

In recent years, generative AI has achieved remarkable progress, especially in text-to-video (T2V) diffusion models. These models demonstrate strong capabilities, producing diverse video content from textual prompts. However, generating physically realistic fire scenes remains a significant challenge. Flames and smoke are highly dynamic, nonlinear, and uncertain, making it difficult for current T2V models to ensure both realism and temporal consistency. Traditional approaches to fire video generation face clear limitations. Data-driven T2V methods often rely on large-scale datasets, but collecting fire scene videos is costly and difficult. Training-free T2V methods, on the other hand, often suffer from poor temporal consistency and coarse visual details.

Recent advances [1, 2] indicate that the quality of prompts has a substantial impact on generative performance. Prior studies have shown that carefully expanded prompts can significantly enhance color richness, motion plausibility, and cross-frame consistency in video generation. This insight suggests that prompt design may be a promising solution for fire scene generation, where data resources are limited.

Thus, we have built T2VFire following the framework of Xue et al. [2] This framework requires no additional training and can be integrated into multiple existing video generation models to enhance the realism and coherence of the generated videos. Specifically, we first use retrieval-augmented generation (RAG) to provide multimodal large language models with fire-domain knowledge, improving physical plausibility and semantic relevance of generated videos. Second, we have designed a prompt expansion mechanism based on first-frame guidance and user feedback to ensure that scene layouts and dynamic details match user intent. Finally, we apply a multi-round prompt refinement process to continuously correct physical violations and semantic deviations.

1 Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. E-mail: H. Zheng, hongtao.zheng@connect.polyu.hk; X. Huang, xy.huang@polyu.edu.hk (✉).

2 Widemount Dynamics Tech Limited, Hong Kong SAR, China.

Manuscript received: 2024-10-12; accepted: 2025-09-09

2 Proposed method

2.1 Stage 1: RAG with GPT-4o

The first stage introduces an RAG to dynamically provide GPT-4o with highly relevant fire-domain knowledge. See Fig. 1(a). The process starts when a user issues a query. The input is first semantically encoded using a text embedding model. The system then matches the encoded query with a pre-built vector database through semantic similarity. These sources can include structured and unstructured documents from authoritative platforms such as Elsevier, Springer, Google Scholar, IEEE, arXiv, and Baidu. However, when selecting database sources, it is important to consider ethical norms and commercial authorization restrictions to ensure compliance and legality of retrieval and usage. Optionally, to ensure freshness and relevance, the database can be dynamically expanded by using keywords from the query to fetch and embed new passages. This step enables knowledge augmentation to be customized in real time for the user's query. Finally, the most relevant text segments retrieved from the database are provided to GPT-4o as additional context. It is important to emphasize that Stage 1 does not allow GPT-4o to autonomously retrieve information from the web. Its retrieval scope is limited to our internally constructed database.

2.2 Stage 2: Establishing the tone of the video

The second stage generates a high-quality first-frame image as the main tone for subsequent video generation. This image is physically plausible and

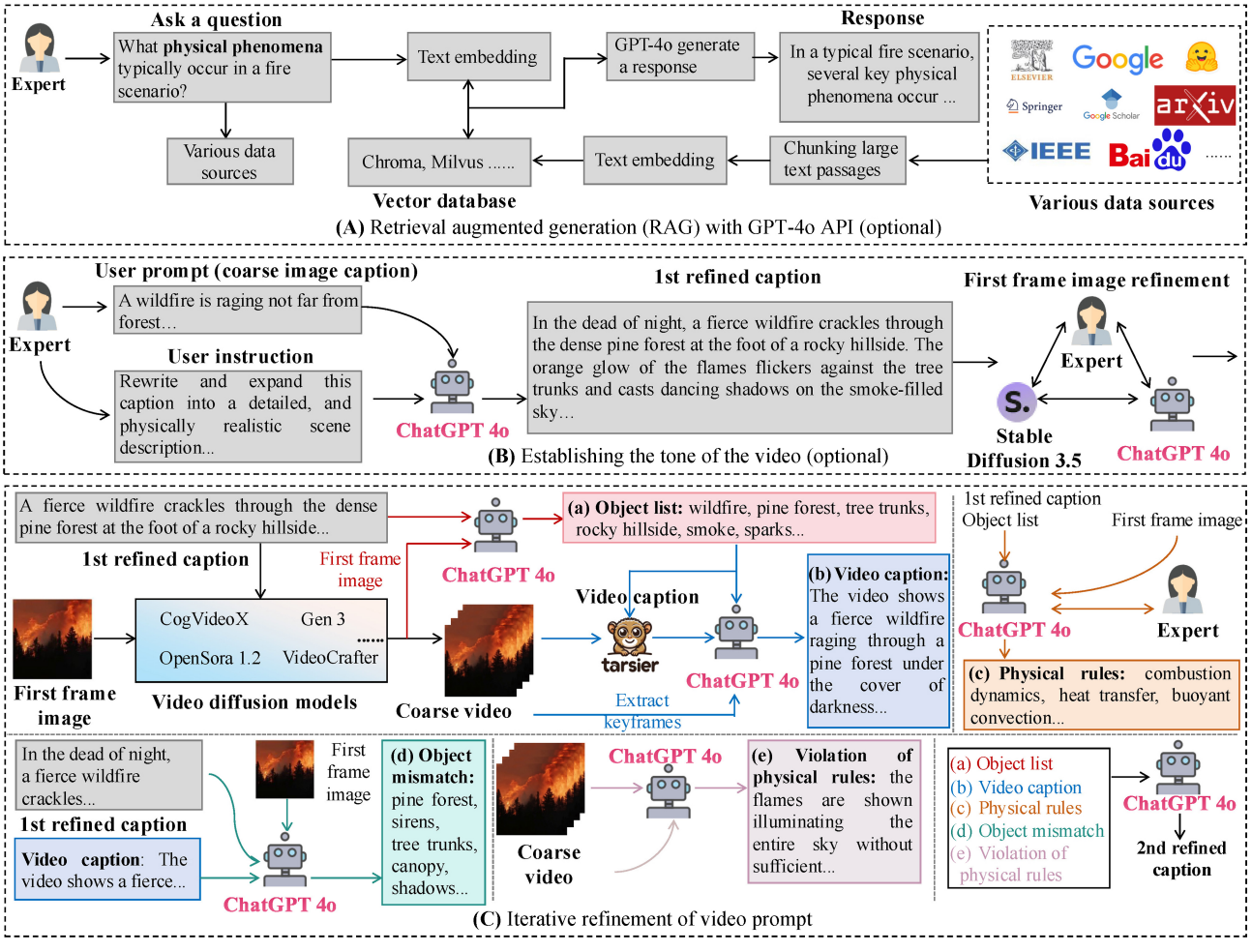

Fig. 1 Inference pipeline of the T2VFire framework. Its three stages are: (A) GPT-4o-based retrieval-augmented generation. Given a user query, semantically relevant text is retrieved from a dynamically expandable vector database and provided to GPT-4o for domain-specific response generation. (B) First-frame generation to set video tone. A coarse text prompt T_0 is expanded by GPT-4o into a detailed scene description T_1, which is used by a text-to-image model (Stable Diffusion 3.5) to generate the first video frame. (C) Iterative prompt refinement. Note that Stage 3 can function independently even without Stage 1 and Stage 2.

aligned with the user's intent, as shown in Fig. 1(b). Since users often find it difficult to describe complex fire scenes in detail, prompts may lack sufficient detail and physical constraints. To address this, we use GPT-4o to expand the coarse input T_0 into a more structured, detailed, and physically consistent prompt T_1, covering elements such as scene composition (foreground and background) and dynamic cues (fire spreading and smoke drifting with the wind). If necessary, a user feedback loop can be introduced to iteratively refine both the prompt and the generated image. By default, Stable Diffusion 3.5 [3] is used to generate the first-frame image in this stage. If the video generation model in Stage 3 is image to video (I2V), the expanded prompt and first-frame image are jointly used as input, providing a stable and coherent foundation for multi-round prompt refinement. If the model in Stage 3 is T2V, this stage by default only performs prompt expansion without generating an image.

2.3 Stage 3: Iterative refinement of video prompt

In the third stage, the expanded text description and the first-frame image are used to guide a GPT-4o through multiple rounds of iterative reasoning. The goal is to generate increasingly refined text prompts that improve the physical realism and semantic alignment of the generated videos. In each iteration, we divide the prompt refinement task into a two-level subproblem structure, following the approach proposed by Xue et al. [2], using five parallel subproblems and one final subproblem, as shown in Fig. 1(c).

The five parallel subproblems each focus on a specific aspect of prompt–video consistency. Each of these subproblems is addressed through local chain-of-thought (CoT) reasoning, as introduced by Xue et al. [2], where GPT-4o conducts step-by-step analysis within each module to ensure local accuracy and clarity of the prompt content. The task instructions used to prompt GPT-4o to perform CoT reasoning for each subtask are shown in Fig. 2(a). We begin by inputting the current prompt T_1 into a pretrained video generation model $\mathcal{G}_{\text{vid}}$, which produces a coarse fire video $V_1 = \mathcal{G}_{\text{vid}}(T_1)$, where V_1 represents the coarse video clip generated based on the user's intent and serves as the input for all subsequent analysis.

We now describe the five parallel subtasks, and the final task in turn.

2.3.1 Object element extraction

In the first step, we aim to infer a structured list of objects that should be present in the generated video, based on the user's intent and the visual context. To achieve this, we feed the prompt T_1 and the first frame of the coarse video $V_1[0]$ into GPT-4o. The model performs cross-modal reasoning to summarize the expected object list $\mathcal{O}_1^{\text{text}} = \mathcal{L}(T_1, V_1[0])$, where $\mathcal{O}_1^{\text{text}}$ represents the set of key visual elements that should be present in fire scene according to the user's description and the initial visual output.

2.3.2 Video captioning

To better align the generated video content with the user's intent, we apply semantic correction and refinement to the initial video caption.

First, a coarse caption C_1^{raw} is generated using an external video captioning model $\mathcal{C}$ (e.g., Tarsier2 [4]) on the coarse video V_1, $C_1^{\text{raw}} = \mathcal{C}(V_1)$.

Then, GPT-4o receives three types of input: (i) the key object list $\mathcal{O}_1^{\text{text}}$ extracted in Task 1, (ii) the raw caption C_1^{raw}, and (iii) a set of representative keyframes selected from the coarse video. GPT-4o performs structured analysis and semantic adjustments based on both the object list and the visual content. Specifically:

- For any object in the list that is not mentioned in the caption, it appends a factual statement.
- For any content in the caption unrelated to the object list, it moves the corresponding sentence to the end.

It also reorders the caption sentences to improve semantic coherence.

2.3.3 Physical rule extraction

Physical realism is a key criterion in evaluating the quality of fire scenes. In this step, GPT-4o is prompted to act as a physics expert and fire science expert. The model receives as input the user prompt T_1, the object list $\mathcal{O}_1^{\text{text}}$ extracted in Task 1, and the first frame of the video $V_1[0]$. Based on these inputs, GPT-4o performs structured reasoning to infer the set of implicit physical rules expected in the described scene, $\mathcal{P}_1^{\text{text}} = \mathcal{L}(T_1, \mathcal{O}_1^{\text{text}}, V_1[0])$, where $\mathcal{P}_1^{\text{text}}$ denotes a natural language description of physical behavior. We also provide GPT-4o with prior instructions to pay special attention to common physical inconsistencies in fire scene video generation, such as contradictory spreading directions of flames and smoke, to enhance its ability to detect and correct physical inconsistencies.

Step 1: Object Element Extraction

Task instruction: You are a scene analyst and fire science expert. Your task is to deeply understand the user-uploaded image (used as the first frame of a video) and identify the key visual elements in the image, using the user-provided prompt as a reference. Ultimately, please provide a detailed list of object categories that are critical to the realism of the fire scene. I**n-context examples** are provided below for reference. Please complete the current task based on this framework.

Step 2: Video Captioning

Task instruction: You are a professional video captioner and fire science expert. Your task is to match and reorder the caption content according to the object list:
Missing objects: For any object in the list that is not mentioned in the caption, append a factual sentence at the end.
Irrelevant content: For any object mentioned in the caption but not found in the object list, move the corresponding sentence to the end of the caption.

Step 3: Physical Rule Extraction

Task instruction: You are a physics expert and fire science expert. Your task is to infer the set of physical rules implicitly expected in the described fire scene. This includes, but is not limited to, the natural behaviors of fire and smoke (e.g., rising, spreading, drifting with wind), as well as physical properties or motions related to other mentioned objects. Use clear natural language to describe these rules. Do not use formulas or symbols and aim to reflect a strong understanding of physical realism in fire-related scenarios.

Step 4: Semantic Mismatch Detection

Task instruction: You are a semantic understanding and fire science expert. Your task is to identify object-level semantic mismatches between the user intent and the generated video. Follow the instructions below: 1. For each object or key element mentioned in the user prompt, check if it appears clearly in the first frame or is mentioned in the caption. 2. If an object is missing, not visible, or wrongly described, mark it as a mismatch. 3. Use clear and natural language to describe the mismatch. For example: “Smoke is mentioned in the prompt but not visible in the frame.” “The prompt describes flames moving uphill, but there is no slope or upward motion in the image.”

Step 5: Physical Violation Summarization

Task instruction: You are a physics expert and fire science specialist. Your task is to determine whether the video contains any violations of physical rules and explain them clearly in natural language. Follow these steps: 1. Read the physical rules and understand the expected behaviors, such as “flames should rise” or “smoke should drift with the wind.” 2. Check the keyframe sequence for any visual evidence that contradicts these rules. 3. For each confirmed violation, describe it clearly in natural language. Focus on physical inconsistencies in the temporal dynamics of fire and smoke, such as discontinuous movement, unrealistic directions, or missing elements.

(a) Five parallel sub-tasks

You are a prompt engineering expert and fire safety specialist using a diffusion model to generate fire scene videos from input prompts. Your task is to refine the current prompt to ensure the generated video better simulates physical realism and aligns more accurately with user intent. You will be provided with an object list, video caption, physical rules, mismatches between the current video and prompt, and a summary of physical violations for reference. Your output should be a rewritten prompt that describes the expected video content in natural language. Focus on correcting parts of the scene that previously exhibited severe physical inconsistencies or inaccurate representations of key elements. Do not mention any technical or physical terminology explicitly. The output must not exceed 150 English words.

(b) One final sub-task

Fig. 2 Example of the task instruction-driven mechanism in Stage 3. (a) Task instructions used to guide GPT-4o in completing five parallel subproblems, including visual element identification, video captioning, physical rule extraction, semantic mismatch detection, and physical violation identification. Each subproblem is solved using local chain-of-thought reasoning. (b) Task instructions for the final subproblem, where GPT-4o integrates the analysis results from the five parallel modules and applies global step-back reasoning to refine the user prompt.

2.3.4 *Semantic mismatch detection*

This step aims to identify object-level mismatches between the user's intent and the actual visual and textual content of the generated video. GPT-4o is prompted to act as a semantic reasoning and fire science expert. The model receives three inputs: the user prompt T_1, the refined video caption C_1, and the first frame $V_1[0]$. It performs cross-modal reasoning to detect inconsistencies between the expected and actual scene content, $\mathcal{M}_1 = \mathcal{L}(T_1, C_1, V_1[0])$.

Specifically, GPT-4o checks whether each key object or element mentioned in the prompt is (i) visible in the image and (ii) referenced in the caption. If an object is missing, not clearly represented, or inaccurately described, it is marked as a mismatch. Each mismatch is expressed in natural language.

2.3.5 *Physical violation detection*

This step aims to detect whether the generated video exhibits any violations of basic physical rules, especially in fire-related behavior, such as flame motion and smoke dynamics.

The model is provided with two inputs: (i) a sequence of keyframes $\mathcal{F}_1 = \{V_1[t_0], V_1[t_1], \ldots, V_1[t_n]\}$, sampled at 1 frame/s from the generated video V_1, and (ii) the set of inferred physical rules $\mathcal{P}_1^{\text{text}}$ from the third task. GPT-4o conducts rule-based evaluation by comparing each physical rule in $\mathcal{P}_1^{\text{text}}$ with the visual evidence in the keyframe sequence. It checks whether expected physical behaviors, such as "flames rising" or "smoke drifting with wind", are visually present and temporally consistent. Violations are identified when the visual content contradicts the expected physical dynamics.

The result is a set of natural language descriptions for each confirmed violation, $\mathcal{R}_1 = \mathcal{L}(\mathcal{P}_1^{\text{text}}, \mathcal{F}_1)$, where $\mathcal{R}_1$ represents the list of detected physical inconsistencies.

2.3.6 *Final subproblem*

After completing the five parallel subproblems, GPT-4o performs one final subproblem to integrate and reflect on the overall feedback, as shown in Fig. 2(b). This task is guided by a global step-back reasoning strategy, as introduced by Xue et al. [2], in which the model reviews the outputs of the local modules from a higher-level perspective to identify cross-module inconsistencies, missing abstractions, or conflicting cues.

Specifically, the outputs from the five subproblems are collected into a structured feedback set, $\Phi_1 = \{\mathcal{O}_1^{\text{text}}, \mathcal{O}_1^{\text{video}}, C_1, \mathcal{P}_1^{\text{text}}, \mathcal{M}_1, \mathcal{R}_1\}$.

GPT-4o receives this feedback along with the current prompt T_1, and is instructed to reason across different abstraction levels (e.g., object, scene, event) and modules (e.g., semantics, physics). Using a global step-back reasoning process, the model produces an improved prompt T_2, $T_2 = \mathcal{L}(T_1, \Phi_1)$.

This new prompt serves as the input for the next iteration, guiding the video generation model toward better alignment with user intent. The iterative loop continues until the generated video achieves the desired level of physical plausibility and semantic consistency. The number of iterations is set to 3.

3 Results

3.1 Qualitative evaluation

We present a qualitative comparison with current state-of-the-art methods in Figs. 3 and 4. As shown in Fig. 3, CogVideoX-5B often produces fire and smoke spreading in opposite directions, or generates fire and smoke without cause. These results lack physical properties and dynamic consistency. In contrast, T2VFire uses physical constraints and a self-iterative mechanism to remove unreasonable motion patterns and visual artifacts, producing more realistic fire videos. Figure 4 further shows that even the most advanced closed-source model still generates clear physical errors. Examples include objects appearing out of nowhere or fire and smoke spreading in implausible directions. These errors often occur in regions not mentioned in the prompt. The advantage of T2VFire is that it acts as a "physical observer", continuously monitoring and correcting potential inconsistencies during generation.

3.2 Quantitative evaluation

Since there are no unified and reliable metrics to comprehensively evaluate the physical plausibility of fire scenes, we use three common video quality metrics in our experiments: motion smoothness, temporal flickering [5], and CLIP score [6]. Although these metrics are not specifically designed for fire-related videos, they can indirectly reflect the dynamic consistency, temporal stability, and semantic alignment of the generated videos. Table 1 compares T2VFire with several state-of-the-art video generation models, such as CogVideoX, CogvideoX-I2V [7],

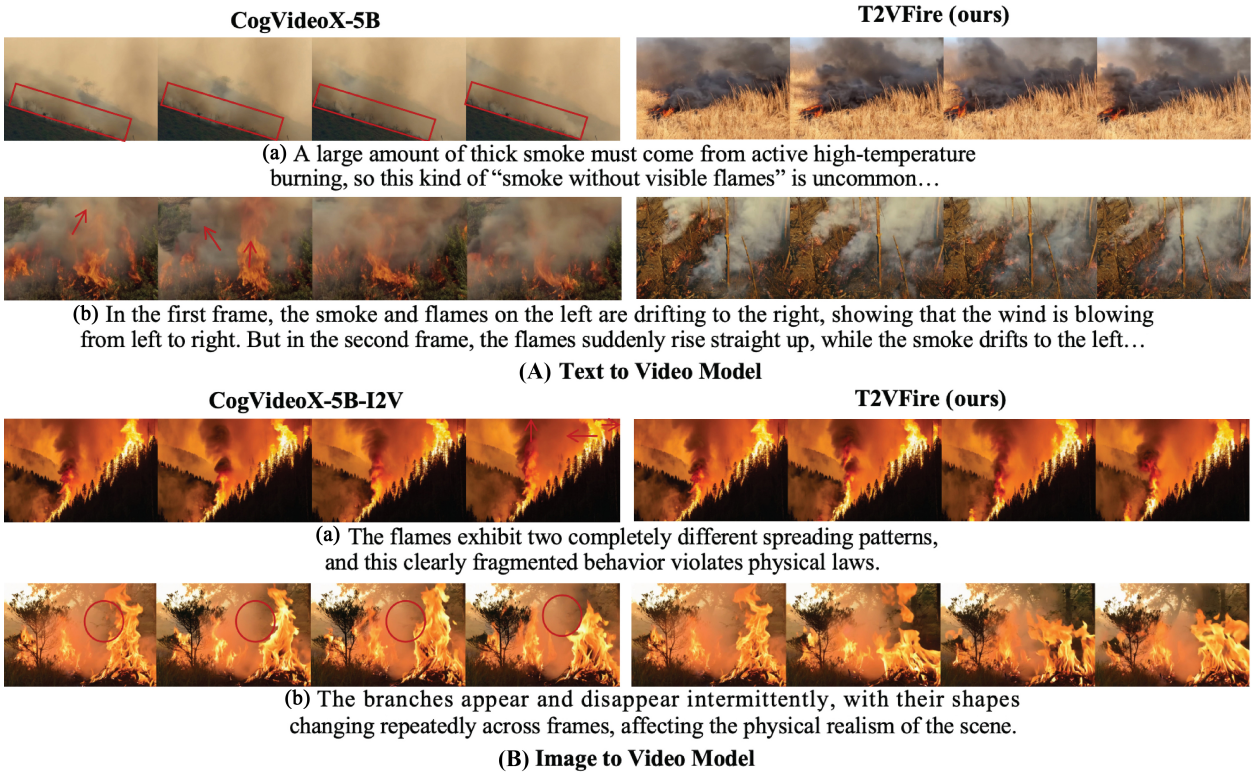

Fig. 3 Comparison between the open-source T2V model CogVideoX-5B, I2V model CogVideoX-5B-I2V, and our method T2VFire. Each video subtitle shows the specific improvements in physical consistency achieved by T2VFire over the earlier video generation model.

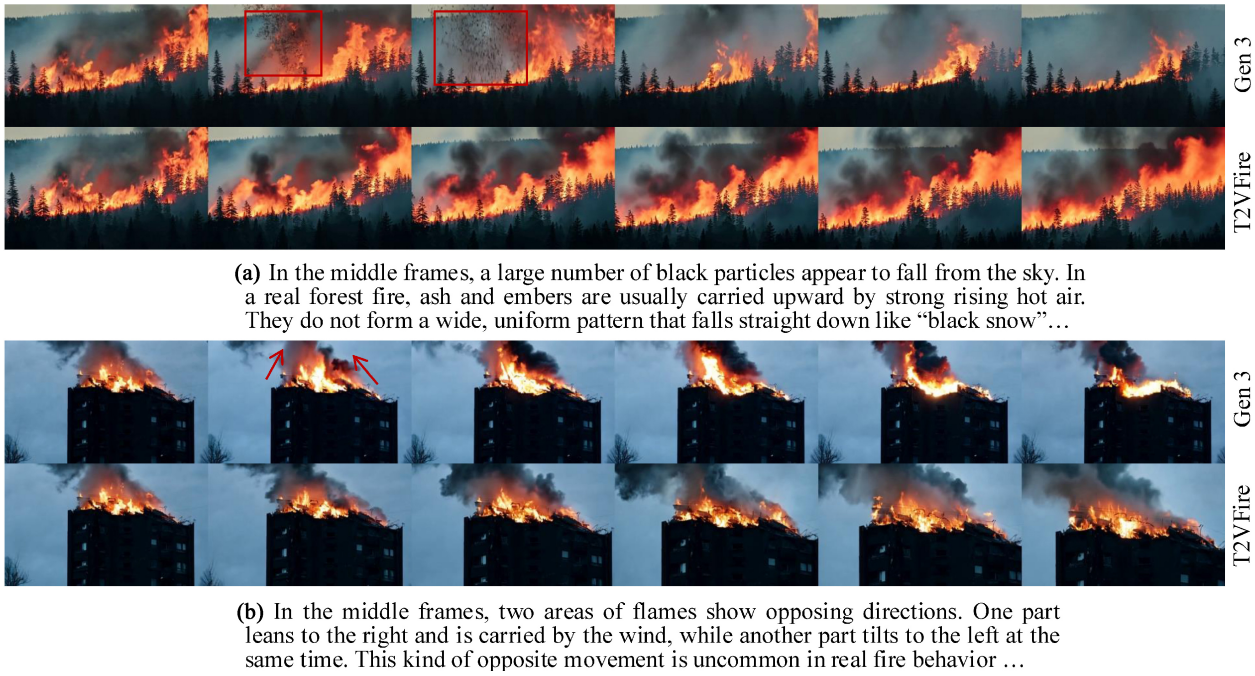

Fig. 4 Existing commercial closed-source video diffusion models, such as Runway Gen-3, cannot generate physically plausible motion. Each video subtitle shows the specific improvements in physical consistency achieved by T2VFire over the earlier video generation model.

and LTX-Video-I2V [8]. We also report results after one iteration (T2VFire (1)) and two iterations (T2VFire (2)) to evaluate the effectiveness of the self-iterative mechanism. Overall, T2VFire outperforms the baselines in most cases, and performance continues to improve with more iterations.

Table 1 Video quality comparison of different T2V models and I2V models after 1 or 2 rounds of iteration in T2VFire

Method	Motion↑	CLIP-Score↑	Flickering↑
T2V model			
CogVideoX	0.972	0.270	0.965
T2VFire (1)	0.977	0.278	0.964
T2VFire (2)	0.983	0.282	0.972
I2V model			
LTX-Video-I2V	0.965	0.260	0.969
T2VFire (1)	0.971	0.266	0.967
T2VFire (2)	0.977	0.279	0.974
CogvideoX-I2V	0.970	0.266	0.972
T2VFire (1)	0.974	0.273	0.970
T2VFire (2)	0.986	0.280	0.976

However, it should be emphasized that the utility of these quantitative metrics has mainly been validated on general video generation tasks. For fire scenes, which involve nonlinear dynamics and strong lighting variations, these metrics cannot directly reflect physical consistency. Therefore, the quantitative results in this experiment serve only as indications, while our main evaluation focuses on qualitative analysis to more comprehensively assess the physical plausibility and visual realism of the generated videos.

4 Conclusions

This paper has presented T2VFire, a training-free T2V framework for generating physically consistent fire scenes. The method integrates a pretrained multimodal language model with an iterative optimization mechanism, effectively improving the physical realism and temporal coherence of fire scene videos produced by general video generation models.

Nevertheless, several limitations should be noted. The performance of T2VFire still relies heavily on the reasoning capability of GPT-4o and the paired diffusion model, and multiple iterations inevitably increase generation time. In addition, GPT-4o may fail to detect some unconventional physical violations, especially those involving complex temporal dynamics or fluid-mechanical behavior. The system's reliance on prompt design and retrieval quality also introduces potential instability in semantic and physical consistency. Meanwhile, the current quantitative evaluation mainly adopts general video generation metrics, which are not sufficient to deeply reflect physical accuracy. Future work will focus on developing physics-aware evaluation metrics and extending the framework to broader physical domains such as fluid motion, explosions, and smoke simulation.

Nevertheless, overall, T2VFire significantly enhances the physical consistency of existing video generation models without additional training, and provides a promising foundation for intelligent firefighting simulation, disaster visualization, and digital-twin-based safety systems.

Acknowledgements

XH thanks the financial support from RGC Theme-based Research Scheme (T22-505/19-N).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Shin, J.; Tang, C.; Mohati, T.; Nayebi, M.; Wang, S.; Hemmati, H. Prompt engineering or fine-tuning: An empirical assessment of LLMs for code. *arXiv preprint* arXiv:2310.10508, 2023.

[2] Xue, Q.; Yin, X.; Yang, B.; Gao, W. PhyT2V: LLM-guided iterative self-refinement for physics-grounded text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18826–18836, 2025.

[3] Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint* arXiv:2403.03206, 2024.

[4] Yuan, L.; Wang, J.; Sun, H.; Zhang, Y.; Lin, Y. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint* arXiv:2501.07888, 2025.

[5] Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 21807–21818, 2024.

[6] Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; Shan, Y. EvalCrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22139–22149, 2024.

[7] Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint* arXiv:2408.06072, 2024.

[8] HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; et al. LTX-video: Realtime video latent diffusion. *arXiv preprint* arXiv:2501.00103, 2024.

Hongtao Zheng is currently conducting research related to artificial intelligence and fire under the supervision of Prof. Xinyan Huang at Hong Kong Polytechnic University, where he will continue his Ph.D. studies. His research focuses on generative AI and its applications to fire dynamics.

Xinyan Huang is an associate professor at Hong Kong Polytechnic University. He received his Ph.D. degree from Imperial College London and was a postdoc at the University of California, Berkeley. His research focuses on the use of AI to solve engineering problems. He has co-authored over 200 peer-reviewed papers and supervised over 40 Ph.D. students and postdocs. He is an associate editor of *Fire Technology* and *the International Journal of Wildland Fire*, and is an editorial board member of other journals. He is a board member of the Int. Assoc. Fire Safety Science and the Int. Assoc. Wildland Fire.

To submit a manuscript, please go to `https://jcvm.org`.