P-Flow: Prompting Visual Effects Generation

Rui Zhao, Mike Zheng Shou*
Show Lab, National University of Singapore

Abstract

Recent advancements in video generation models have significantly improved their ability to follow text prompts. However, the customization of dynamic visual effects, defined as temporally evolving and appearance-driven visual phenomena like object crushing or explosion, remains underexplored. Prior works on motion customization or control mainly focus on low-level motions of the subject or camera, which can be guided using explicit control signals such as motion trajectories. In contrast, dynamic visual effects involve higher-level semantics that are more naturally suited for control via text prompts. However, it is hard and time-consuming for humans to craft a single prompt that accurately specifies these effects, as they require complex temporal reasoning and iterative refinement over time. To address this challenge, we propose P-Flow, a novel training-free framework for customizing dynamic visual effects in video generation without modifying the underlying model. By leveraging the semantic and temporal reasoning capabilities of vision-language models, P-Flow performs test-time prompt optimization, refining prompts based on the discrepancy between the visual effects of the reference video and the generated output. Through iterative refinement, the prompts evolve to better induce the desired dynamic effect in novel scenes. Experiments demonstrate that P-Flow achieves high-fidelity and diverse visual effect customization and outperforms other models on both text-to-video and image-to-video generation tasks. Code is available at https://github.com/showlab/P-Flow.

Figure 1. It is hard for humans to craft text prompts that precisely control video generation models to generate desired visual effects across diverse scenes, while P-Flow automatically refines prompts to achieve consistent and realistic visual effects.

1. Introduction

Recent advancements in video generation models have significantly enhanced their ability to produce visually compelling content guided by text instructions [1, 32, 69]. These models excel at generating videos that align with high-level semantic descriptions, enabling applications in creative storytelling, virtual environments, and visual design [10, 17, 20, 60, 86]. However, specifying nuanced, temporally evolving phenomena, such as *dynamic visual effects* (e.g., object explosion, crushing), remains a challenge. Unlike low-level motion control [76, 91], which can be guided by explicit trajectories, dynamic visual effects require higher-level semantic understanding and temporal coherence, making them difficult to capture with explicit conditions.

While such effects are naturally suited for control via text prompts due to their semantic richness, crafting prompts that accurately describe dynamic visual effects is inherently complex. Users must articulate both the semantic characteristics and temporal evolution of the effect, often requiring

*Corresponding Author.

iterative refinement and complex temporal reasoning. For instance, applying a reference explosion effect to a new scene, such as a meteor crashing into the moon, requires preserving the dynamics and timing of the effect while adapting it to a completely different visual and semantic context, as shown in Fig. 1. Manual prompt engineering for such tasks is time-consuming and often yields suboptimal results.

Prior works on video customization have primarily focused on low-level motion control, such as guiding subject or camera motion using trajectories or spatial paths [23, 76]. While effective for explicit motion tasks, these methods are ill-suited for high-level semantic effects that lack clear motion trajectories. Alternative approaches that fine-tune video generation models for specific effects require extensive computational resources and lack generalizability across diverse effects [44]. In contrast, a training-free paradigm that leverages the powerful abilities of foundational generation models would offer a flexible and user-friendly solution for effect customization.

To address these challenges, we propose P-Flow, a novel training-free framework that customizes dynamic visual effects in video generation by treating text prompts as optimization variables. Rather than updating the generation model itself, P-Flow performs test-time prompt optimization, leveraging the semantic and temporal reasoning capabilities of vision-language models (VLMs) to iteratively refine prompts and bridge the gap between generated video and reference visual effects. To make this optimization both effective and stable, we introduce two key strategies. First, we introduce a noise prior that emphasizes temporally salient dynamics in the reference effect to guide stable optimization, while incorporating stochastic noise to maintain diversity and exploration during prompt refinement. Second, we incorporate a lightweight historical context mechanism that maintains past optimization trajectories, enabling more consistent and coherent refinement across iterations. Together, these designs ensure that prompts evolve meaningfully over time, achieving high-fidelity visual effects customization.

The experimental results validate the effectiveness and generality of P-Flow in enabling high-fidelity and diverse visual effect generation across both image-to-video and text-to-video generation settings. Without any model fine-tuning, P-Flow achieves state-of-the-art performance in key metrics such as FID-VID [68], FVD [6], and Dynamic Degree [28], and is strongly preferred in human evaluations. Compared to the training-based baseline constrained by fixed-length supervision and training dataset biases, our test-time optimization approach fully captures the temporal evolution of effects and better adapts to diverse scenes. These findings demonstrate the potential of P-Flow as a plug-and-play solution for dynamic visual effect generation.

Our code will be fully open-sourced. The main contributions are summarized as follows: (1) We propose P-Flow, a training-free framework that customizes dynamic visual effects in video generation by optimizing text prompts at test time. It supports both text-to-video and image-to-video generation. (2) We introduce a novel prompt optimization paradigm guided by VLM, enhanced with a noise prior to stabilize learning while preserving diversity, and a lightweight historical context mechanism to ensure optimization coherence. (3) Extensive experiments demonstrate the state-of-the-art performance of P-Flow across metrics and human evaluations.

2. Related Works

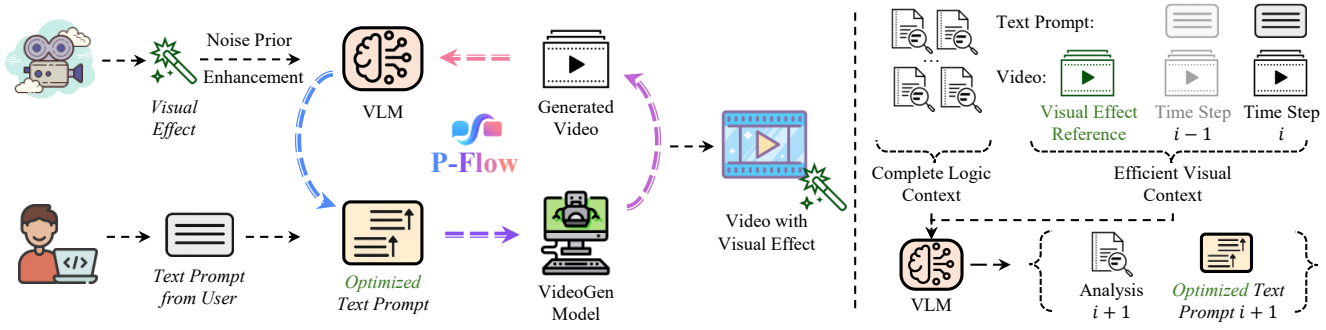
2.1. Video Generation Model

Recent generation models demonstrate their powerful abilities in generating diverse and high- fidelity contents [24, 25, 46, 64, 88, 92, 94]. Where video generation approaches are largely based on diffusion models [7, 8, 26, 37, 38, 52, 71, 73, 74, 87, 97], which generate videos by denoising Gaussian noise through architectures such as 3D U-Net [61] or transformer-based DiT [56]. More recently, flow matching models [31, 40, 43, 47] have emerged as a scalable and efficient alternative, directly learning a velocity field to map noise to data without iterative denoising, and have shown superior quality on both realistic and diverse video generation tasks [32, 69]. And a growing number of open-source video generation models [21, 39, 57, 84, 84, 96] have recently been released, offering diverse architectures and capabilities for both text and image conditioned video generation. The increasing fidelity and prompt-following ability of open-sourced SOTA models provide a promising foundation for prompt-based video generation and optimization.

2.2. Motion Customization and Control

Motion customization methods [91] extend subject and style customization [13, 19, 34, 63, 65, 77] to the temporal domain by enabling control over motion dynamics. DreamVideo [78] and LAMP [80] learn motion patterns or adapters to customize both appearance and motion. Other methods [29, 51, 59, 75, 83] further explore disentangled or reference-guided motion generation. In parallel, controllable video generation aims to ensure the generation results align with the given explicit control signals, such as depth maps, human pose, optical flows, etc. [5, 23, 48, 49, 58, 76, 82, 89, 90, 95]. These methods mainly address low-level motion using explicit priors or training-based control modules [12, 72].

In contrast, dynamic visual effects involve higher-level semantics and remain underexplored. A concurrent work, VFX Creator [44], adds control branches for visual effect generation but is limited to image-to-video generation and requires separate training for each different type of visual effect. Our method offers a flexible, training-free solution



(a) Framework. (b) Historical Trajectory Maintenance.

Figure 2. Overview of the proposed P-Flow framework.

applicable to text-to-video and image-to-video models.

2.3. Vision Language Models for Generation

Recent advances in large language models [2, 3, 9, 55, 67] have significantly enhanced the capabilities of vision-language models (VLMs) [4, 11, 36, 41, 66, 98], enabling them to perform semantic and temporal reasoning over visual content. These models have been increasingly used to evaluate or guide generation [14–16, 18, 22, 27, 30, 35, 45, 50, 53, 54, 79, 81, 85], with Gecko [79] demonstrating their effectiveness in assessing fine-grained generation quality across diverse attributes. Some recent work, such as EvolveDirector [93] and VideoAlign [42], explored the use of VLM to train and optimize generation models to align human preferences. However, applying VLMs for test-time optimization in video generation remains largely unexplored. Our work leverages VLMs not only for evaluation but also as optimization tools to bridge the semantic gap between text prompts and complex visual effects.

3. Method

3.1. Problem Formulation

Given a reference video V_{ref} showing a dynamic visual effect and an initial text prompt P_0 describing a novel scene or subject, our objective is to generate a video V_{gen} that exhibits the same visual effect as V_{ref} while adhering to the semantic content specified by P_0.

In the image-to-video generation task, the generation is additionally conditioned on a source image I that provides detailed spatial structure or appearance of the scene. In this case, the generated video is given by $V_{\text{gen}} = \mathcal{G}(P^*, I, \eta)$, where $\mathcal{G}$ is a pre-trained video generation model and η denotes latent noise. For simplicity, and unless otherwise stated, we omit I in the formulation to unify notation across both text-to-video (T2V) and image-to-video (I2V) scenarios.

Formally, we aim to optimize a text prompt P^* such that the generated video $V_{\text{gen}} = \mathcal{G}(P^*, \eta)$ minimizes the discrepancy $\mathcal{D}(V_{\text{gen}}, V_{\text{ref}})$ in terms of the semantic and temporal characteristics of the visual effect.

3.2. Framework Overview

The P-Flow framework operates in a training-free manner, optimizing the text prompt at test time without modifying the underlying video generation model. The method comprises three core components: (1) noise prior enhancement to initialize the latent noise for stable and diverse video sampling, (2) test-time prompt optimization using a VLM to iteratively refine the prompt, and (3) historical trajectory maintenance to guide the refinement decisions of VLM. The process is iterative, generating videos, evaluating their alignment with the reference effect, and refining the prompt until a maximum number of iterations is reached.

3.3. Noise Prior Enhancement

We found that the initial latent noise η used in video generation significantly influences optimization stability and output diversity. Completely random noise results in inconsistent visual effects across text prompt optimization iterations, hindering convergence, while fixed noise limits exploration, leading to suboptimal solutions. To address this, we propose a noise prior enhancement strategy that balances stability and exploration through flow matching inversion, temporal noise isolation, and noise blending.

First, we extract the latent noise corresponding to V_{ref} via flow matching inversion [33, 62, 70]. In flow matching, the generative model defines a continuous-time ordinary differential equation (ODE)

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta\big(x_t, t; P\big), \tag{1}$$

which transports noise η at $t = 0$ to the data x_T at $t = T$. To invert this process, we integrate the same vector field backward in time starting from $x_T = V_{\text{ref}}$ with its corresponding reference prompt P_{ref}:

$$\eta_{\text{inv}} = x_0 = x_T - \int_0^T v_\theta\big(x_t, t; P_{\text{ref}}\big)\,\mathrm{d}t. \tag{2}$$

By construction, this ensures $\mathcal{G}(P_{\text{ref}}, \eta_{\text{inv}}) \approx V_{\text{ref}}$, where η_{inv} captures both the dynamic visual effect and appearance-specific attributes (e.g., textures or background elements) that are orthogonal to the visual effect itself.

To isolate the motion-related temporal components from the inverted noise $\eta_{\text{inv}} \in \mathbb{R}^{C\times F\times H\times W}$, where C is the number of latent channels, F is the number of frames, and H, W are spatial dimensions, we apply a two-stage SVD-based projection. First, we reshape η_{inv} into a matrix $\mathbf{N}_s \in \mathbb{R}^{(C\cdot F)\times(H\cdot W)}$ and compute its singular value decomposition:

$$\mathbf{N}_s = \mathbf{U}_s\mathbf{\Sigma}_s\mathbf{V}_s^\top. \tag{3}$$

To suppress appearance-specific spatial variations, we adaptively determine the number of leading components k_s to remove by ensuring the retained energy satisfies

$$\frac{\sum_{i=k_s+1}^{r_s}\sigma_i^2}{\sum_{i=1}^{r_s}\sigma_i^2} \geq \rho_s, \tag{4}$$

where $r_s = \text{rank}(\mathbf{N}_s)$. We set the top k_s singular values in $\mathbf{\Sigma}_s$ to zero and reconstruct the spatially-filtered tensor as

$$\eta_{\text{spatial}} = \text{reshape}\left(\mathbf{U}_s\mathbf{\Sigma}_s'\mathbf{V}_s^\top, [C, F, H, W]\right). \tag{5}$$

Next, η_{spatial} is reshaped along the temporal axis into $\mathbf{N}_m \in \mathbb{R}^{(C\cdot H\cdot W)\times F}$ and do SVD project again, and we retain the top k_m components such that

$$\frac{\sum_{i=1}^{k_m}\sigma_i'^2}{\sum_{i=1}^{r_m}\sigma_i'^2} \geq \rho_m. \tag{6}$$

The final projected noise $\eta_{\text{temporal}} \in \mathbb{R}^{C\times F\times H\times W}$ preserves dominant motion information while suppressing static and appearance-dependent details.

Finally, to ensure exploratory diversity, we blend η_{temporal} with random noise $\eta_{\text{new}} \sim \mathcal{N}(0, I)$:

$$\eta = \sqrt{\alpha}\cdot\eta_{\text{temporal}} + \sqrt{1-\alpha}\cdot\eta_{\text{new}}, \tag{7}$$

where α controls the influence of the motion-preserving noise. This blended noise η is used to sample the video $V_{\text{gen}} = \mathcal{G}(P_i, \eta)$ at iteration i.

3.4. Test-Time Prompt Optimization

At each iteration i, we generate a video V_{gen}^i using the current prompt P_i and the enhanced noise η as

$$V_{\text{gen}}^i = \mathcal{G}(P_i, \eta), \tag{8}$$

where $\mathcal{G}$ is the video generation model. To assess the alignment between the generated visual effects and those in the reference video V_{ref}, we construct a composite video by vertically stacking V_{ref}, the previously generated video (if available), and V_{gen}^i. The composite video V_{comb} is preprocessed to ensure consistent resolution and frame rate, enabling direct visual comparison across inputs.

A VLM is employed to analyze differences between V_{gen}^i and V_{ref}, focusing on motion dynamics and visual effects, while explicitly ignoring variations in appearance or identity. Based on this analysis, the VLM performs prompt refinement to guide the next generation toward better reproducing the target visual effects:

$$P_{i+1} = \mathcal{M}(V_{\text{comb}}, P_i, \mathcal{H}; P_0) \tag{9}$$

Here, $\mathcal{M}(\cdot)$ denotes the VLM structured refinement function, which takes as input the reference and generated video pair V_{comb}, the current prompt P_i, the historical trajectory of optimization, detailed in Sec. 3.5, and the original content constraints from P_0. The output is an updated prompt P_{i+1}, where only effect-related descriptions are modified, preserving the original subject and environment.

The VLM is instructed to return a structured JSON object containing detailed analysis and the revised prompt P_{i+1}. This iterative process enables fine-grained control over visual effect fidelity through prompt optimization. The full procedure is presented as pseudocode in the Appendix.

3.5. Historical Trajectory Maintenance

To enhance the reasoning and optimization capabilities of the VLM, we maintain a historical trajectory

$$\mathcal{H} = \{(V_i, P_i, A_i)\}_{i=0}^{i_{\max}-1}, \tag{10}$$

where V_i, P_i, and A_i denote the generated video, the corresponding prompt, and the VLM analysis at iteration i. This trajectory provides context for prompt refinement, allowing the VLM to identify effective optimization patterns and avoid redundant changes. For example, if previous iterations have consistently increased the intensity of a desired visual effect, the VLM may favor similar refinements in subsequent steps.

However, storing the full sequence of previously generated videos introduces considerable computational overhead, especially as video inputs consume large amounts of visual tokens in the VLM. To address this, we adopt a memory-efficient strategy: only the reference video, the generated video from the previous iteration, and the current generated video are included in the visual input to the VLM. This selection maintains the most relevant temporal context while significantly reducing token length.

Meanwhile, to preserve long-term memory and optimization history, we retain all text prompts $\{P_i\}$ and VLM analyses $\{A_i\}$ across iterations in $\mathcal{H}$. Since language tokens are much more compact than visual tokens, this design provides a good trade-off between efficiency and contextual richness, enabling the VLM to reason over past refinements while operating within practical computational constraints.

Table 1. Quantitative comparisons for both image-to-video and text-to-video generation settings. Note that VFX Creator supports only image-to-video generation.

	Image-to-Video			Text-to-Video			Overall		
	FID-VID ↓	FVD ↓	Dyn. Degree ↑	FID-VID ↓	FVD ↓	Dyn. Degree ↑	FID-VID ↓	FVD ↓	Dyn. Degree ↑
Wan 2.1 [69]	34.62	994.70	0.33	42.32	1535.43	0.28	38.47	1265.07	0.31
HunyuanVideo [32]	36.53	1169.9	0.51	38.35	1362.35	0.34	37.44	1266.13	0.43
VFX Creator (*Training-Based*) [44]	29.92	**752.95**	0.63	–	–	–	–	–	–
Wan 2.1 + HF (*Training-Free*)	33.12	989.42	0.54	42.21	1632.10	0.59	75.33	1310.76	0.57
HunyuanVideo + HF (*Training-Free*)	33.85	1035.49	0.70	36.54	1266.79	0.62	35.20	1151.14	0.66
P-Flow (Ours, *Training-Free*)	**29.32**	784.51	**0.94**	**32.93**	**980.75**	**0.87**	**31.13**	**882.63**	**0.91**

3.6. Implementation Details

We use the pre-trained Wan 2.1 14B video generation models [69] for both text-to-video and image-to-video tasks, producing videos at the resolution of 480×832 with 81 frames. For image-to-video generation, the aspect ratio is adaptively adjusted to be the same as the input image. Text prompt optimization is performed using the Gemini 1.5 Pro vision-language model. The blending weight is fixed to $\alpha = 0.001$, and the optimization process is run for $i_{\max} = 10$ iterations. All experiments are conducted on an NVIDIA A100 GPU cluster. Video generation is performed with 8-GPU distributed inference, taking approximately 69 seconds per video and consuming around 40 GB of GPU memory per card. In each optimization iteration, besides the video generation, 1.2 seconds are used to construct the input for VLM, and 16.3 seconds are spent on prompt refinement via VLM inference. Structured instructions for VLM and more details are provided in the Appendix.

4. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of P-Flow in customizing dynamic visual effects for video generation. The evaluation spans a diverse set of visual effects and includes both objective metrics and subjective human preference studies. We compare P-Flow with recent state-of-the-art methods through quantitative results in Sec. 4.2 and qualitative visualizations in Sec. 4.3. In addition, we perform an ablation study in Sec. 4.4 to analyze the contribution of key components. Experimental settings are detailed in Sec. 4.1.

4.1. Experimental Setup

Dataset: The experiments are conducted on the Open-VFX dataset [44]. This benchmark comprises 675 high-quality videos sourced from commercial platforms, where each video lasts approximately 5 seconds at 24 fps. These videos span 15 diverse categories of dynamic visual effects, such as explode, deflate, and squish, offering rich visual diversity and temporal dynamics. Additionally, 245 reference images are provided for the image-to-video generation task, covering both single and multi-object scenes. We sample reference videos from its training set and test images from its test set.

Metrics: To assess the visual effect fidelity and dynamism of generated videos, we adopt three standard metrics following prior work [44]: FID-VID [68]: Fréchet Inception Distance adapted for videos, measuring distributional similarity between generated and ground-truth videos. FVD [6]: Fréchet Video Distance, which captures temporal coherence and realism based on a 3D video feature extractor. Dynamic Degree [28]: Quantifies the degree of motion or visual transformation across frames to reflect effect intensity and temporal variation.

In addition, we conduct a human evaluation using a pairwise comparison protocol, where 15 annotators are asked to choose the better video between two candidates in terms of visual effect fidelity. For each generation task, we sampled 100 samples with 15 different types of visual effects from each model.

Baselines: We compare P-Flow against the foundational state-of-the-art video generation models, Wan 2.1 [69] and HunyuanVideo [32], as well as a prior specialized model, VFX Creator [44], which is specifically designed for visual effect learning. On the Open-VFX dataset, VFX Creator is trained with a separate LoRA version for each type of visual effect. All baselines are used with their publicly released checkpoints and configurations. We additionally include a human feedback (HF) mode for Wan 2.1 and HunyuanVideo, where the text prompt is manually revised once, based on the generated results, to improve the visual alignment with the given visual effect references.

4.2. Quantitative Results

As shown in Table 1, our proposed method P-Flow achieves superior or highly competitive performance across all metrics in both image-to-video and text-to-video generation tasks. P-Flow is built upon Wan 2.1 in our experiments.

Specifically, P-Flow outperforms strong foundational video generation models such as Wan 2.1 and HunyuanVideo across all three metrics in both generation settings. Notably, our method achieves this without any fine-tuning or modification of the foundational model parameters, demonstrating the effectiveness of our test-time optimization framework.

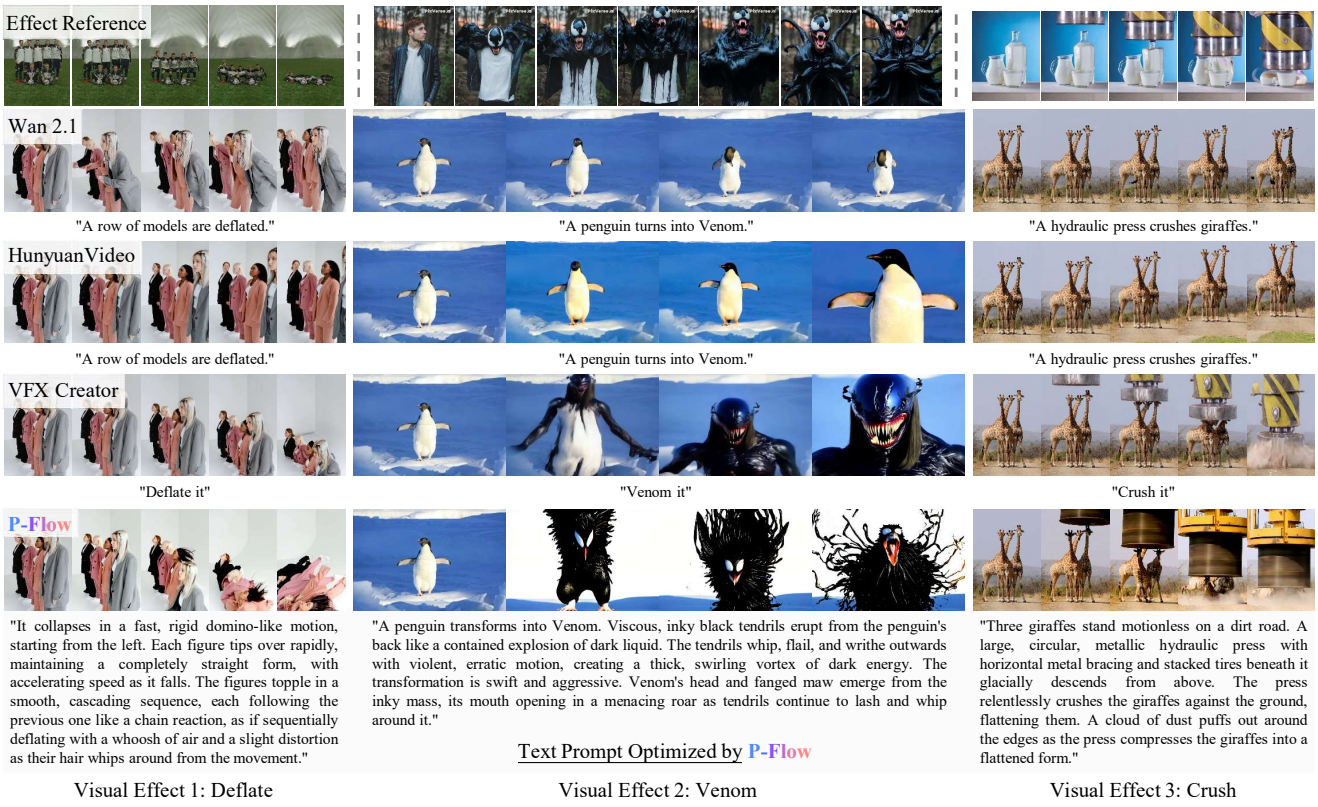


Figure 3. Qualitative comparison on image-to-video generation with different visual effects. The prompts shown beneath each row represent the actual input to each model. As VFX Creator is optimized for short phrase inputs, such prompts are provided to ensure a fair and consistent evaluation.

Table 2. Human evaluation results (%) comparing P-Flow against baseline models in Image-to-Video (I2V) and Text-to-Video (T2V) generation. Note: Model order was randomized during evaluation.

Model 1	Preference (Model 1 V.S. Model 2)	Model 2
P-Flow-I2V	**80%** V.S. 20%	Wan 2.1-I2V
P-Flow-I2V	**84%** V.S. 16%	HunyuanVideo-I2V
P-Flow-I2V	**58%** V.S. 42%	VFX Creator
P-Flow-T2V	**75%** V.S. 25%	Wan 2.1-T2V
P-Flow-T2V	**81%** V.S. 19%	HunyuanVideo-T2V

This validates our design philosophy of treating the video generator as a black box while still enabling high-quality visual effect generation through adaptive, input-specific optimization.

Compared to the training-based method VFX Creator, which is trained on the Open-VFX dataset and involves dedicated architectural designs, our method achieves comparable results in FID-VID and FVD, while significantly outperforming it in Dynamic Degree. This highlights the strength of our method in generating videos with more salient and temporally coherent motion, which is essential for visual effects generation.

Moreover, it is worth noting that VFX Creator does not support text-to-video generation, and its trained LoRA weights are tightly coupled with specific architectures. In contrast, P-Flow is training-free, modular, and model-agnostic, supporting both image-to-video and text-to-video tasks. Achieving such generalization and performance without any training overhead underscores the practicality and robustness of our framework.

A pairwise human preference study is conducted to compare the visual effect fidelity of P-Flow with others. As shown in Table 2, the results demonstrate that P-Flow consistently outperforms existing models in both settings, reflecting its superiority in visual effect generation.

4.3. Qualitative Results

As shown in Fig. 3 and Fig. 4, our proposed P-Flow demonstrates clear advantages in generating high-quality and controllable visual effects. It is worth mentioning that, for the P-Flow, there are no constraints on the resolution or length of the reference video. This greatly reduces the barrier for users to adopt our method, allowing them to freely choose reference clips of any duration or resolution.

The pre-trained strong foundational models, Wan 2.1 and HunyuanVideo, fail to produce the desired effects using plain text prompts, which highlights the insufficiency of

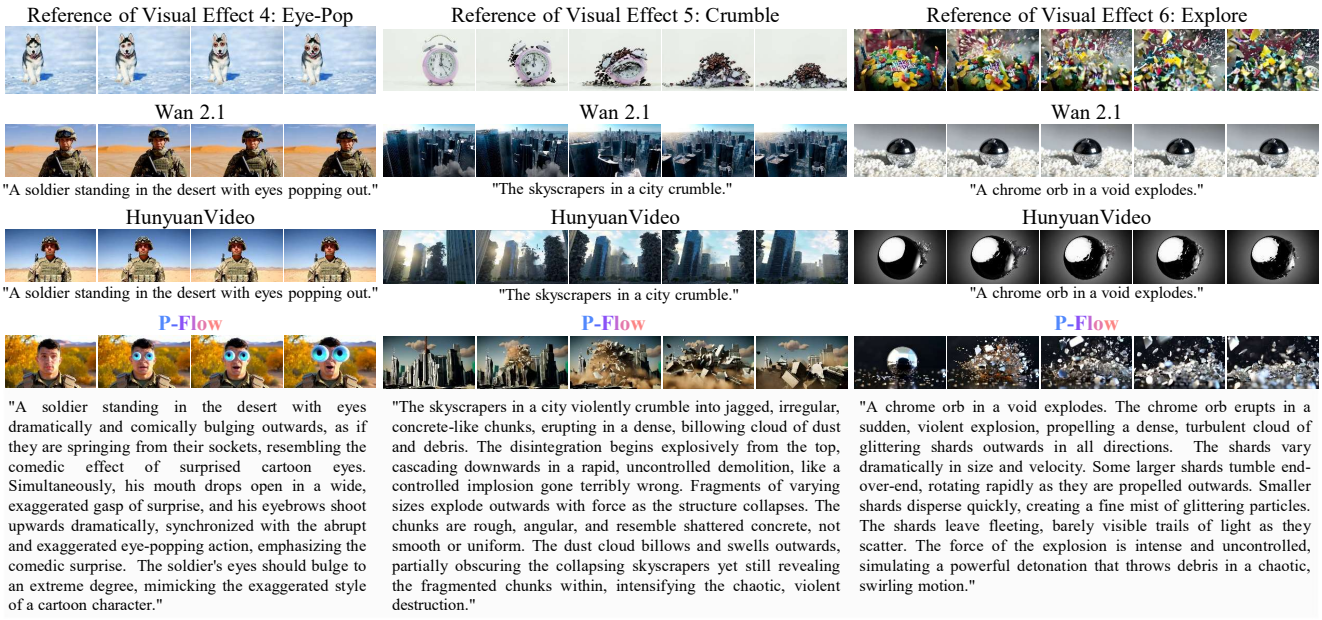


Figure 4. Qualitative comparison on image-to-video generation with different visual effects. The prompts shown beneath each row represent the actual input to each model.

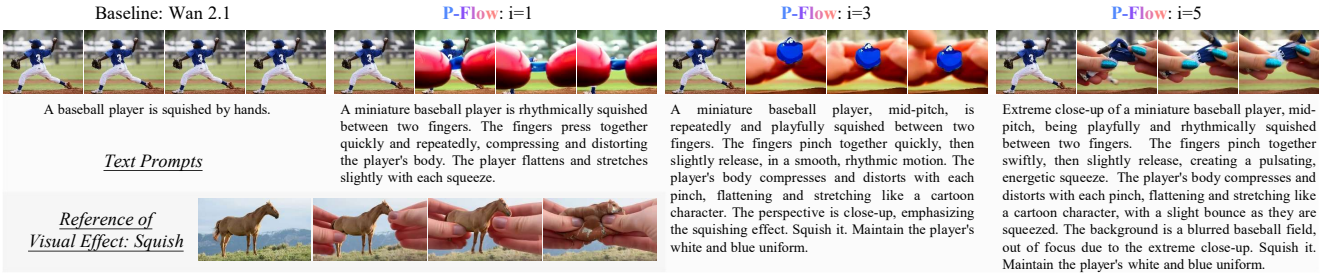


Figure 5. Optimization trajectory of P-Flow. Starting from a simple base prompt, P-Flow iteratively refines the text prompt based on the visual feedback (we showcase the iteration $1 \rightarrow 3 \rightarrow 5$), leading to progressively more accurate alignment between the generated video and the target "squish" visual effect.

generic prompts in steering these models toward specific visual goals.

In comparison, the training-based model, VFX Creator, exhibits relatively stronger ability in capturing visual effects. Nevertheless, it also suffers from inherent limitations imposed by its fixed-length training regime. For example, in the *Visual Effect 1: Deflation*, the synthesized sequence terminates before the visual transformation completes. This truncation is attributed to that all training samples are forcibly trimmed to a fixed length, which may lead to cutting out some parts of the visual effects. P-Flow, in contrast, imposes no such constraint. The full reference video can be encoded by the VLM, allowing the dynamic evolution of the effect to be fully captured and reflected in the optimized prompt, thereby avoiding truncation-related failures.

In addition, the training-based method may also encode dataset-specific biases. For example, in *Visual Effect 2: Venom*, the second frame generated by VFX Creator includes a humanoid body structure, likely due to bias in the training data toward human-centric subjects. These artifacts reveal the limited generalization capacity of training-based models under distribution shifts. Our method, by optimizing the prompt at inference time based on the input image and reference video, naturally avoids such artifacts, accurately preserving subject-specific attributes from the input image while incorporating the desired visual effect from references. The results in Fig. 4 further demonstrate the superiority of our method on the text-to-video generation task.

Optimization Trajectory. We visualize the prompt optimization trajectory of P-Flow in Fig. 5. Given a reference video containing the desired visual effect, P-Flow gradually optimizes the text prompt to guide the generation towards similar dynamics in a novel scene.

Table 3. Ablation study of P-Flow on both image-to-video and text-to-video generation.

Modules			Image-to-Video			Text-to-Video			Overall		
Noise-Enhance	Logic-Context	Visual-Context (i-1)	FID-VID ↓	FVD ↓	Dyn. Degree ↑	FID-VID ↓	FVD ↓	Dyn. Degree ↑	FID-VID ↓	FVD ↓	Dyn. Degree ↑
✗	✗	✗	33.42	1089.23	0.64	39.85	1321.70	0.61	36.64	1205.47	0.63
✓	✗	✗	32.80	961.78	0.69	36.74	1182.42	0.66	34.77	1072.10	0.68
✓	✗	✓	30.45	861.52	0.83	34.05	1044.67	0.78	32.25	953.10	0.81
✓	✓	✓	**29.32**	**784.51**	**0.94**	**32.93**	**980.75**	**0.87**	**31.13**	**882.63**	**0.91**

4.4. Ablation Study

We conduct the ablation study to investigate the effectiveness of each component in our framework, including the Noise-Enhance module, the Visual-Context (i-1), and the Logic-Context modules. Results are summarized in Table 3 under both image-to-video and text-to-video settings. Specifically, Visual-Context (i-1) refers to incorporating the previously generated video frame at time step i-1 as visual context for the current generation.

It is shown that even without incorporating any of the three ablation components, the performance of P-Flow already surpasses the strong foundational model, Wan 2.1 [69], in terms of Dynamic Degree. This demonstrates that text prompt optimization alone, without any tuning or additional temporal modules, can significantly enhance the temporal dynamics.

As shown in Table 3, each module contributes incrementally to performance. Adding the Noise-Enhance component leads to improvements because it can stabilize the optimization of our framework. Introducing short-term context through the Visual-Context module brings further gains by offering visual insights for VLM to better analyze the influence of the text prompt and further optimize it. Finally, we incorporate the Logic-Context module, which provides long-range semantic analysis context derived from the entire optimization trajectory. This allows the prompt to maintain high-level coherence and effect progression over time. Notably, by decoupling long-term logic context from short-term visual context, our method avoids the computational overhead of processing long visual sequences, while still benefiting from both temporal scales.

4.5. Hyperparameter Analysis

Our noise prior enhancement strategy involves hyperparameters that control the trade-off between optimization stability and generation diversity. We conduct a parameter study on mage-to-video generation to analyze their effects, and summarize the results as follows.

Energy Thresholds for SVD Projection. We introduce two energy thresholds, ρ_s and ρ_m, to determine the number of principal components retained or suppressed during the two-stage SVD-based projection:

- Spatial energy threshold (ρ_s): Controls the suppression of appearance-related spatial details, e.g., textures, tone, and background patterns.
- Temporal energy threshold (ρ_m): Determines the amount of motion-relevant temporal variation to retain.

Table 4. Analysis of noise prior enhancement components.

Setting	FID-VID ↓	FVD ↓	Dyn. Degree ↑
w/o SVD Projection ($\rho_s = 0.0, \rho_m = 1.0$)	33.25	1052.80	0.58
Random Noise Only ($\alpha = 0.0$)	32.74	923.67	0.73
Enhanced Noise ($\alpha = 0.001, \rho_s = 0.1, \rho_m = 0.9$)	29.32	**784.51**	**0.94**
Enhanced Noise ($\alpha = 0.01, \rho_s = 0.1, \rho_m = 0.9$)	**29.21**	803.35	0.88

When setting $\rho_s = 0$, i.e., without spatial suppression, the model retains unwanted appearance priors from the reference video, resulting in degraded visual quality, i.e. FID-VID: 33.25 and FVD: 1052.80, as shown in Table 4. On the other hand, we empirically observed that setting ρ_s too high (> 0.5) overly suppresses useful priors, leading to diminished impact of the enhanced noise. A moderate value $\rho_s = 0.1$ achieves the best balance.

For the temporal energy threshold, we set $\rho_m = 0.9$ to retain most of the motion-relevant information. As shown in Table 4, these settings help preserve temporal dynamics and achieve a good dynamic score as 0.94.

Blending Coefficient α. We further study the impact of the blending coefficient $\alpha \in [0, 1]$, which controls the mixture of preserved temporal noise and fresh random noise.

As shown in Table 4, using only random noise $\alpha = 0$ achieves limited performance because of the unstable optimization process. A suitable coefficient $\alpha = 0.001$ leads to significant improvement across all metrics, including FVD and motion dynamics, by preserving key information of motion dynamics while introducing sufficient randomness. Slightly increasing α to 0.01 further improves FID-VID but reduces dynamic scores, reflecting a trade-off between fidelity and motion dynamics. We finally set α to 0.001 to achieve better dynamical generation.

5. Conclusion

We present P-Flow, a training-free framework for customizing dynamic visual effects in video generation through test-time prompt optimization. By leveraging noise prior enhancement and historical trajectory, P-Flow enables stable and coherent effect transfer without model fine-tuning. Extensive experiments demonstrate its strong performance and generality, highlighting P-Flow as a practical framework for generating high-fidelity visual effects at test-time.

6. Acknowledgement

This research is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023).

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3

[5] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 2

[6] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 2, 5

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2

[8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 2

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 1

[11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 3

[12] Weifeng Chen, Jie Wu, Pan Xie, Hefeng Wu, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023. 2

[13] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[14] Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. *arXiv preprint arXiv:2503.20491*, 2025. 3

[15] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[16] Yang Du, Zhuoran Lin, Kaiqiang Song, Biao Wang, Zhicheng Zheng, Tiezheng Ge, Bo Zheng, and Qin Jin. Vc4vg: Optimizing video captions for text-to-video generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1138, 2025. 3

[17] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024. 1

[18] Bingjie Gao, Xinyu Gao, Xiaoxue Wu, Yujie Zhou, Yu Qiao, Li Niu, Xinyuan Chen, and Yaohui Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3173–3183, 2025. 3

[19] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. 2

[20] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025. 1

[21] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2

[22] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023. 3

[23] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations*. 2

[24] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. 2

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 2

[27] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 3

[28] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 5

[29] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. *arXiv preprint arXiv:2312.00845*, 2023. 2

[30] Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian Ge, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18725–18735, 2025. 3

[31] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024. 2

[32] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2, 5

[33] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024. 3

[34] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[35] Seunghun Lee, Jihoon Lee, Chan Ho Bae, Myung-Seok Choi, Ryong Lee, and Sangtae Ahn. Optimizing prompts using in-context few-shot learning for text-to-image generative models. *IEEE Access*, 12:2660–2673, 2024. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[37] Xin Li, Wenqing Chu, Ye Wu, Weihang Yuan, Fanglong Liu, Qi Zhang, Fu Li, Haocheng Feng, Errui Ding, and Jingdong Wang. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 2

[38] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv:2309.17444*, 2023. 2

[39] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 2

[40] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[42] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 3

[43] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 2

[44] Xinyu Liu, Ailing Zeng, Wei Xue, Harry Yang, Wenhan Luo, Qifeng Liu, and Yike Guo. Vfx creator: Animated visual effect generation with controllable diffusion transformer. *arXiv preprint arXiv:2502.05979*, 2025. 2, 5

[45] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[46] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[47] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 2

[48] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv:2304.01186*, 2023. 2

[49] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 2

[50] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana

Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 3

[51] Joanna Materzynska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Customizing motion in text-to-video diffusion models. *arXiv preprint arXiv:2312.04966*, 2023. 2

[52] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023. 2

[53] Wenyi Mo, Tianyu Zhang, Yalong Bai, Bing Su, Ji-Rong Wen, and Qing Yang. Dynamic prompt optimizing for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26627–26636, 2024. 3

[54] Hyelin Nam, Jaemin Kim, Dohun Lee, and Jong Chul Ye. Optical-flow guided prompt optimization for coherent video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7837–7846, 2025. 3

[55] OpenAI. GPT-4 technical report, 2023. 3

[56] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[57] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025. 2

[58] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751*, 2025. 2

[59] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780*, 2024. 2

[60] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. *arXiv preprint arXiv:2501.09781*, 2025. 1

[61] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[62] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024. 3

[63] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[64] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[65] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 2

[66] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[67] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 3

[68] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2, 5

[69] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 2, 5, 8

[70] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 3

[71] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Videofactory: Swap attention in spatiotemporal diffusions for text-to-video generation. *arXiv:2305.10874*, 2023. 2

[72] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2

[73] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Pe der Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Y. Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. 2023. 2

[74] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv:2309.15103*, 2023. 2

[75] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023. 2

[76] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video

generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 2

[77] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[78] Yujie Wei, Shiwei Zhang, Hangjie Yuan, Xiang Wang, Haonan Qiu, Rui Zhao, Yutong Feng, Feng Liu, Zhizhong Huang, Jiaxin Ye, et al. Dreamvideo-2: Zero-shot subject-driven video customization with precise motion control. *arXiv preprint arXiv:2410.13830*, 2024. 2

[79] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024. 3

[80] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 2

[81] Dawei Xiang, Wenyan Xu, Kexin Chu, Tianqi Ding, Zixu Shen, Yiming Zeng, Jianchang Su, and Wei Zhang. Promptsculptor: Multi-agent based text-to-image prompt optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 774–786, 2025. 3

[82] Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *arXiv:2306.00943*, 2023. 2

[83] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. *arXiv preprint arXiv:2402.03162*, 2024. 2

[84] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *ICLR*, 2025. 2

[85] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[86] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025. 1

[87] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18456–18466, 2023. 2

[88] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 2

[89] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[90] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. *arXiv preprint arXiv:2407.21705*, 2024. 2

[91] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023. 1, 2

[92] Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, and Changjie Fan. Zero-shot text-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21013–21023, 2023. 2

[93] Rui Zhao, Hangjie Yuan, Yujie Wei, Shiwei Zhang, Yuchao Gu, Lingmin Ran, Xiang Wang, Jay Zhangjie Wu, David Junhao Zhang, Yingya Zhang, et al. Evolvedirector: Approaching advanced text-to-image generation with large vision-language models. *Advances in Neural Information Processing Systems*, 37:122104–122129, 2024. 3

[94] Rui Zhao, Weijia Mao, and Mike Zheng Shou. Doracycle: Domain-oriented adaptation of unified generative model in multimodal cycles. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2835–2846, 2025. 2

[95] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Unicontrolnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 2

[96] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2

[97] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv:2211.11018*, 2022. 2

[98] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3