

PAI-Studio: Cinematic Video Background Replacement with Camera-Aware Motion

Heyuan Gao^{1,2*}, Bangxun Tang^{1,3*}, Yiren Song^{1,4*†}, Guian Fang^{1,4}, Zijian He¹, Jie Yang¹, and Mike Zheng Shou^{4†}

¹Utopai Studios, USA ²Nanyang Technological University, Singapore

³University of California, Irvine, USA

⁴Show Lab, National University of Singapore, Singapore

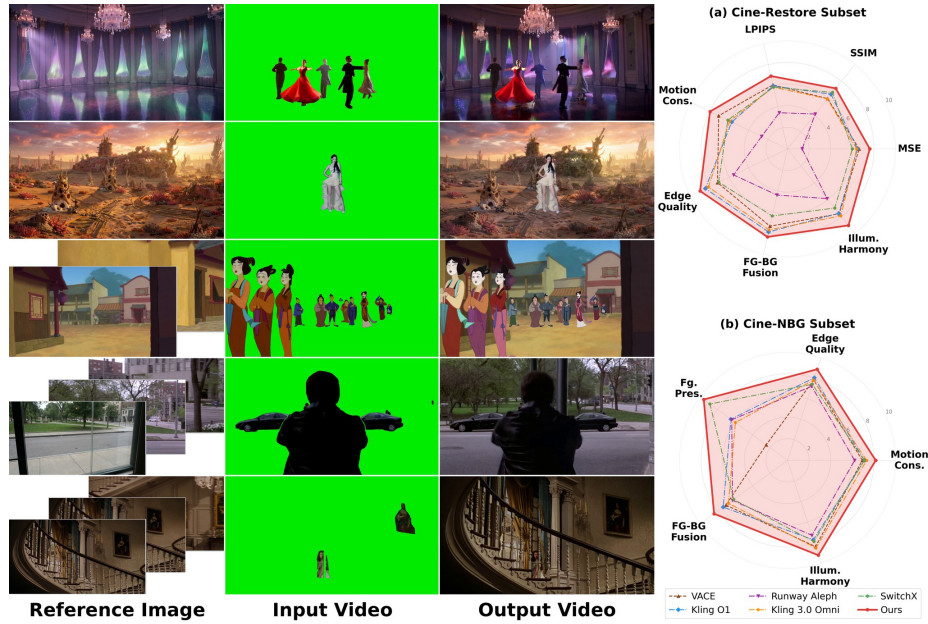


Fig. 1: Given a foreground video and background reference images, our method synthesizes a complete video with camera-aware background motion, foreground consistency, and scene-consistent relighting at 1080P resolution. Readers can click and play the video clips in this figure using [Adobe Acrobat](#).

Abstract. We present PAI-Studio, a new reference-conditioned video synthesis task that addresses a long-standing challenge in cinematic background replacement: generating dynamic backgrounds aligned with foreground motion while preserving foreground identity, matching reference scene appearance, and achieving globally consistent illumination with realistic foreground relighting. Existing open-source systems and commercial APIs cannot simultaneously ensure motion-consistent background generation, high-fidelity foreground relighting and foreground identity

*Equal contribution. ‡Project leader. †Corresponding author.

preservation, often resulting in static backgrounds, inconsistent boundaries, and noticeable compositing artifacts. To bridge this gap, we build upon a Diffusion Transformer video backbone and reformulate the problem as an in-context conditional generation task. Through bidirectional attention, our model jointly captures foreground dynamics and background reference information within a unified architecture. We further construct a 30K-scale dataset sourced from high-quality films and online videos to support this task. Extensive evaluations demonstrate that our method significantly outperforms existing open-source and commercial API solutions. Project page: <https://github.com/showlab/PAI-Studio>.

Keywords: Video Generation · Background Replacement · Diffusion Transformer

1 Introduction

Recent advances in video generation and editing have significantly expanded the possibilities of content creation. However, a critical gap remains in cinematic background replacement: existing systems often fail to generate backgrounds that move consistently with the foreground’s cinematographic motion. In practice, when the foreground involves professional camera movements (e.g., push, pan, tracking shots), synthesized backgrounds frequently appear static, misaligned, or visibly composited. Crucially, most existing methods implicitly rely on the unrealistic assumption of perfect, pixel-accurate foreground mattes. In real-world inference, handling imperfect segmentation is inevitable; flickering edges and broken contours amplify artifacts and reduce realism. Furthermore, methods attempting global harmonization often struggle to maintain the subject’s identity.

Through systematic analysis, we identify that high-quality cinematic background synthesis requires four key properties: **motion consistency** (explicitly modeling implied camera movement and dynamic trajectory), **illumination consistency** (scene-adaptive foreground relighting), **boundary robustness** (tolerating imperfect masks and maintaining coherent transitions), and **foreground consistency** (preserving subject identity and fine textures).

To achieve this, we introduce **PAI-Studio**, a novel video synthesis framework for multi-condition cinematic background replacement. Given a foreground video and one to three background reference images, PAI-Studio generates a complete video satisfying all four properties (Fig. 1). We explicitly support **multi-frame background control** to handle significant background content variations under large camera motion, aligning reference images with designated output frames through positional encoding cloning. At the model level, we build upon an in-context Diffusion Transformer [29] backbone with native bidirectional self-attention. Unlike existing pipelines that stack complex cascaded modules (e.g., dedicated matting networks or relighting adapters), our approach is remarkably elegant and globally reasons over multi-modal conditions without architectural

Table 1: Feature comparison between our method and existing approaches. VACE* denotes our reproduction trained on the same dataset.

Method	1080P	Multi-frame BG Control	Moving Background	Foreground Consistency	Illumination Harmony	Edge Harmony	Background Consistency
VACE*	✓	✗	✓	✗	✓	✓	✗
Kling O1	✓	✓	✓	✗	✓	✓	✓
Runway Gen-4 Aleph	✓	✗	✗	✗	✗	✗	✗
Kling 3.0 Omni	✗	✓	✓	✗	✓	✓	✗
Beeble AI Switchx	✓	✗	✗	✓	✗	✓	✓
Ours	✓	✓	✓	✓	✓	✓	✓

bloat. This *global video generation* strategy rather than local background inpainting naturally reconstructs plausible transitions even with incomplete mask edges, thereby achieving **boundary robustness**.

To empower this framework, we construct **CineStudio**, a 30K-scale dataset explicitly reverse-engineered from real-world user inference scenarios. Crucially, we observe that large-scale film and online video data inherently exhibit causal consistency between foreground motion and background evolution under real camera movements, providing natural supervision for learning motion-aligned background synthesis. Therefore, derived from high-quality films and online videos, we perform shot segmentation and optical-flow-based filtering to retain sequences with significant camera and background motion, which explicitly enhances the **motion consistency** of the synthesized results. Furthermore, to enforce **illumination consistency**, our data curation augments foreground videos with controlled relighting perturbations. The resulting dataset contains diverse cinematographic movements, scene layouts, and lighting conditions, providing a realistic foundation for studying motion-aware background synthesis.

Compared to existing methods (Table 1), our approach provides a more comprehensive solution. Our main contributions are threefold:

- We introduce **PAI-Studio**, a comprehensive video synthesis framework for cinematic background replacement that explicitly requires motion consistency, illumination consistency, and boundary robustness, while supporting multi-frame background control.
- We propose an in-context Diffusion Transformer architecture with bidirectional attention and temporally aligned positional encoding, combining data-driven learning and structured video understanding to achieve motion-aligned background generation and coherent relighting without complex cascaded modules.
- We construct **CineStudio**, a 30K-scale dataset from films and online videos, whose construction is reverse-engineered from inference challenges to support the systematic evaluation of motion-aware background generation.

2 Related Works

2.1 Video Diffusion Models

Video diffusion models have rapidly evolved from 2D U-Nets augmented with temporal layers [3, 7, 10] to native spatiotemporal Diffusion Transformers (DiTs) [8, 29]. DiTs treat video data as patch sequences to capture long-range dependencies, and recent large-scale models [4, 11, 19, 41, 50] have demonstrated exceptional scalability and high-fidelity generation. This unified architecture provides the structural foundation for our framework, enabling scalable integration of diverse latent representations. Furthermore, the sequence-based formulation of DiTs allows heterogeneous signals—foreground videos, reference frames, and noisy latents—to be unified into a single token stream for joint reasoning via full self-attention, eliminating the need for separate control branches. Utilizing this structural property, our framework casts cinematic background replacement as an in-context conditional generation problem.

2.2 Controllable Video Generation

Controllable video generation aims to synthesize temporally coherent videos under various conditionings, progressing from text-to-video [7, 11, 13, 14] to richer spatial controls [5, 16, 25]. For appearance and motion guidance, image-conditioned approaches [3, 45, 46] often struggle with large camera motions, while video-conditioned methods [10, 32, 42] require full reference videos, limiting flexibility. Recent advances also explore process-level supervision [35, 37]. A growing body of work further pursues fine-grained, subject-centric control, including expressive freestyle portrait animation [27], open-domain regional image animation driven by motion prompts [26], spatial-temporal decoupled motion transfer [28], and inpainting-based 4D creation [24]. Despite these efforts, existing methods typically treat backgrounds as static or weakly dynamic and fail to explicitly model camera-aware background motion conditioned on foreground dynamics. In contrast, our work focuses on reference-conditioned background synthesis with explicit temporal alignment and multi-frame background control.

2.3 Video Editing and Background Replacement

Classical video background replacement relies on matting and compositing [21, 22, 39], which requires prepared backgrounds and struggles with moving cameras. Recent diffusion-based video editing methods excel at inpainting [20, 53] and text-driven modifications [2, 12, 31, 33, 44, 48, 49], but typically assume static backgrounds or lack global camera motion modeling. While some works combine matting with generative models [47, 51], they often synthesize backgrounds independently of foreground motion, causing temporal inconsistencies. Concurrently, human-centric editing approaches facilitate interactive garment customization [34], diffusion-transformer-based style transfer [38], and real-time effect generation [36] for identity-preserving manipulation. However, these frameworks are not designed for cinematic background replacement, which requires

the synthesized scene to follow inferred camera trajectories while maintaining the geometric and photometric coherence of the foreground subject. Our method directly addresses this by enforcing temporal alignment via bidirectional attention and position-aware conditioning, enabling camera-aware motion and scene-consistent illumination.

3 Method

3.1 Problem Formulation and Inputs

We study cinematic background replacement as a multi-condition video synthesis task. Given a foreground video $V^{fg} = \{v_t^{fg}\}_{t=1}^T$ and a small set of background reference images $\{I_k^{bg}\}_{k=1}^K$ with $K \in \{1, 2, 3\}$, the goal is to synthesize an output video $V^{out} = \{v_t^{out}\}_{t=1}^T$ that satisfies the following requirements:

(1) Background appearance consistency. The synthesized background should match the reference scene appearance indicated by $\{I_k^{bg}\}$. We explicitly support **multi-frame background control** since under large camera motion, the visible background content can vary substantially over time, making single-frame reference insufficient.

(2) Motion consistency. The background dynamics in V^{out} should be consistent with the cinematographic motion implied by the foreground video V^{fg} , producing coherent camera-aware background motion rather than static or drifting backgrounds.

(3) Foreground fidelity with scene-adaptive relighting. The foreground content in V^{out} should preserve the identity and fine details of V^{fg} , while its appearance (e.g., shading, color tone) should be adaptively adjusted to better match the target scene implied by $\{I_k^{bg}\}$, i.e., realistic foreground relighting and global illumination harmonization.

(4) Robustness to imperfect foreground boundaries. In practice, foreground masks from video segmentation often exhibit temporal instability (e.g., flickering, missing parts that reappear in later frames). V^{out} should remain visually coherent under such imperfections, with plausible boundary completion and temporal consistency enabled by spatiotemporal context.

Optionally, a text prompt c^{txt} can be provided to describe the target scene. We denote all conditions as $\mathcal{C} = (V^{fg}, \{I_k^{bg}\}_{k=1}^K, c^{txt})$.

3.2 Overall Architecture

Our approach builds upon a Diffusion Transformer (DiT) video generation backbone (WAN 2.2-5B) [41], adapted to a conditional video-to-video synthesis setting with multiple heterogeneous conditions. All videos and images are first encoded into a shared latent space using a pretrained video VAE, producing latent tokens for the foreground video, background reference images, and text. An overview of the proposed architecture is illustrated in Fig. 2.

At each diffusion step, we form a unified token sequence and process it with a single-stream Transformer using **bidirectional self-attention**, enabling joint

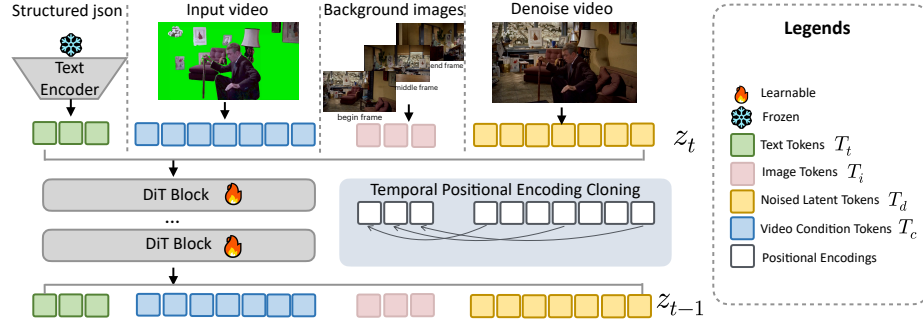


Fig. 2: Overview of the PAI-Studio architecture. Multi-condition inputs—including multiple background reference images, the illumination-perturbed foreground video, a structured prompt JSON, and the denoising video—are encoded into tokens and concatenated. A multi-modal attention model leverages bidirectional attention to model their global correlations. In addition, Temporal Positional Encoding (PE) Cloning is introduced to precisely control the temporal placement of multiple background images in the generated video.

reasoning over foreground dynamics, reference scene appearance, and the evolving denoising state. The denoised latent is decoded back into V^{out} by the VAE decoder. Coupled with our tailored dataset, this native bidirectional attention mechanism efficiently resolves multi-modal consistency and global harmonization challenges.

Full bidirectional self-attention. Let Z^{txt} denote the text tokens (optional), Z^{fg} the foreground video tokens, $\{Z_k^{bg}\}_{k=1}^K$ the background reference tokens, and Z_s the noisy output video tokens at diffusion step s . We concatenate them into a single sequence:

$$X = [Z^{txt} \parallel Z^{fg} \parallel Z_1^{bg} \parallel \dots \parallel Z_K^{bg} \parallel Z_s] \in \mathbb{R}^{N \times d}, \quad (1)$$

where N is the total number of tokens and d is the token dimension. We apply *full* (bidirectional) self-attention on X . For the h -th head,

$$Q_h = XW_h^Q, \quad K_h = XW_h^K, \quad V_h = XW_h^V, \quad (2)$$

$$\text{Attn}_h(X) = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_h}}\right) V_h, \quad (3)$$

and the multi-head output is

$$\text{MHSA}(X) = \text{Concat}(\text{Attn}_1(X), \dots, \text{Attn}_H(X))W^O, \quad (4)$$

where H is the number of heads and $d_h = d/H$.

Structured motion understanding. To strengthen motion and camera-awareness, we leverage a strong video understanding model (Gemini-3) [40] to extract structured cues from videos, including foreground/background motion summaries, camera movement descriptions, and semantic layer descriptions. These structured signals are used for dataset curation and can also be provided at inference time as high-level guidance derived from the input foreground video, improving the stability of motion-aligned synthesis.

3.3 Multi-Frame Background Control via Temporal Positional Encoding Cloning

Single-frame background conditioning is often insufficient under large cinematographic motion, since the visible background content can change substantially across time. We therefore support **multi-frame background control** with $K \in \{1, 2, 3\}$ reference images sampled from designated temporal locations $\{t_k\}_{k=1}^K \subset \{1, \dots, T\}$ (e.g., first/middle/last).

A key challenge is to establish *explicit temporal correspondence* between each background reference I_k^{bg} and a specific output time t_k . We address this with **temporal positional encoding cloning**. Let $\mathbf{p}(t)$ denote the temporal positional encoding used by the DiT for time index t . For each background reference token Z_k^{bg} , we *clone* the temporal positional encoding from the denoising tokens at time t_k :

$$\mathbf{p}(Z_k^{bg}) \leftarrow \mathbf{p}(t_k), \quad k = 1, \dots, K. \quad (5)$$

This design anchors each reference image to a designated output frame and enables controllable, time-aware conditioning.

Moreover, to enforce motion consistency between the input foreground and the synthesized video, we align the temporal positional encodings of corresponding foreground and denoising frames:

$$\mathbf{p}(Z_t^{fg}) \equiv \mathbf{p}(Z_s(t)), \quad t = 1, \dots, T, \quad (6)$$

where $Z_s(t)$ denotes the denoising tokens associated with output time t . Together, Eqs. (5)–(6) provide a simple yet effective mechanism for *temporal controllability* and motion-aligned background synthesis under multi-frame conditioning.

3.4 Illumination Perturbation for Scene-Consistent Relighting

Achieving illumination consistency requires the model to infer foreground lighting from the target scene rather than copying the input foreground illumination. We adopt a simple but effective strategy during data construction: we intentionally apply **relighting and color perturbations** to the foreground video to form $\tilde{V}^{fg}$. This breaks the direct illumination correspondence between the input foreground and the target video, and forces the model to recover appropriate foreground lighting from background/environment cues.

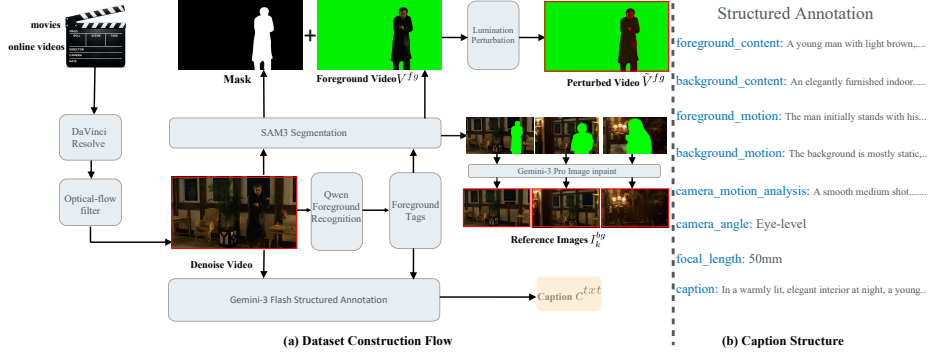


Fig. 3: Overview of the CineStudio data construction pipeline.

Concretely, the perturbation operator $\mathcal{R}(\cdot)$ applies global photometric augmentations, including HSV-channel jittering, contrast adjustment, color temperature shifts, and gamma correction:

$$\tilde{V}^{fg} = \mathcal{R}(V^{fg}). \quad (7)$$

We then train the model with $\tilde{V}^{fg}$ as the conditioning foreground while keeping V^{out} as the target. As a result, the model learns to produce foreground appearances that are consistent with the reference scene illumination, enabling coherent relighting at inference time.

3.5 Dataset Construction

The core philosophy of our data construction pipeline is to explicitly reverse-engineer the training data based on the challenges encountered in real-world user inference. In practical virtual studio scenarios, users consistently struggle with mismatched lighting, imperfect green-screen matting, and the need for temporal coherence. Therefore, rather than collecting ideal synthetic data, we construct our 30K-scale dataset, **CineStudio**, to directly reflect these difficulties: we purposefully inject controlled relighting perturbations, and leverage the natural cinematographic motion of real-world videos. Derived from high-quality films and online videos, each sample contains $(\tilde{V}^{fg}, \{I_k^{bg}\}_{k=1}^K, c^{txt}, V^{out})$, where V^{out} is the original video clip. The overall data construction process is illustrated in Fig. 3.

Shot segmentation and motion filtering. We first perform shot segmentation and compute optical-flow statistics to filter out clips with weak camera/background motion, retaining sequences with clear cinematographic dynamics.

Foreground-background separation and alignment. We perform foreground recognition using Qwen3-VL [1] to obtain a short tag describing the dominant foreground concept. This tag is fed to an advanced video segmentation model

(SAM3) [6] to segment and track the foreground over time, producing temporally aligned masks. These masks are then used to extract the foreground onto a green screen, yielding the final foreground video. The resulting masks may still contain realistic imperfections such as boundary flickering, small holes, or missing regions.

Multi-frame background references. We sample $K \in \{1, 2, 3\}$ background reference frames from different temporal locations (e.g., begin/middle/end), remove the foreground content, and employ Gemini-3 Pro Image [40] to inpaint the missing regions, yielding clean background references $\{I_k^{bg}\}$.

Foreground relighting augmentation. We generate relit variants $\tilde{V}^{fg}$ using the perturbation operator $\mathcal{R}(\cdot)$ (Eq. (7)) to encourage scene-consistent relighting. This augmentation explicitly simulates the inference-time challenge where foreground subjects must dynamically adapt to entirely novel and unpredictable background lighting conditions.

Structured annotation with Gemini-3. Finally, we use Gemini-3 [40] to generate structured JSON annotations for each clip, including foreground/background semantic descriptions, foreground/background motion, camera motion analysis, as well as camera angle and focal length. We also produce a concise caption that summarizes and unifies these fields into a single natural-language description. These annotations are used for dataset curation and for providing high-level motion/camera guidance at inference time.

4 Experiments

4.1 Experimental Setup

Setting. We evaluate on the proposed CineStudio dataset (approximately 30K clips with diverse motions, layouts, and lighting). Each clip contains a foreground video, a target video, and 1-3 background references (defaulting to the **begin, middle, and end** source frames). Videos are generated at 1080P (1920×1056), 24 FPS, and 97 frames ($\sim 4s$). We finetune via LoRA [15] (rank 80) and AdamW [23] (LR 5×10^{-5}) in BF16. Training takes approximately 6,000 steps on 8 H200 GPUs (batch size 1/GPU). Baselines use optimal settings.

Baseline Methods. We compare our method with both open-source and commercial video editing systems that support background replacement or scene synthesis.

(1) Video-editing Baseline. To ensure fair comparison under the same setting, we first exclude models lacking IV2V support. While we surveyed numerous IV2V models, most focus on style transfer or foreground object editing and perform poorly on green-screen background replacement, making them unsuitable as direct baselines (due to space limits, visual results of these baselines are provided

in the **supplementary material**). Therefore, we solely adopt VACE [17] as our open-source baseline. We finetune the VACE-1.3B model on our CineStudio dataset using the same training data and resolution settings as our method.

(2) Commercial APIs. We evaluate Kling O1, Kling 3.0 Omni, Runway Gen-4 Aleph (denoted as Runway Aleph), and Beeble AI Switchx (denoted as Switchx), which provide background replacement or generative scene synthesis functionality. These systems are evaluated under their default API configurations.

Benchmark. To evaluate the proposed method, we introduce a comprehensive benchmark, **Cine-Benchmark**, structured into two distinct subsets to decouple the evaluation of context fidelity and out-of-distribution generalization:

(1) Cine-Restore Subset focuses on *Existing-Frame Background Replacement*. It consists of 150 videos randomly sampled from our **CineStudio** dataset (disjoint from training). In this subset, the background reference frames are clean plates extracted and inpainted from the original source video. This simulates a closed-loop background completion task where the model must faithfully restore the original cinematic context under new lighting conditions.

(2) Cine-NBG (Novel Background Generation) Subset evaluates the model’s synthesis and generalization capabilities in new environments. This *Novel Background Replacement* subset comprises 50 videos curated from public sources (DAVIS 2017 [30] and IVEBench [9]). Unlike Cine-Restore, the background reference frames in Cine-NBG are synthesized by modifying the original background using Gemini to create diverse, out-of-distribution styles and layouts. This subset tests the model’s robustness under real-world camera motions and its ability to generate plausible new backgrounds. Since no ground-truth exists for these generated scenes, we evaluate performance using Gemini-based perceptual scores.

Evaluation Metrics. Following the objectives in Sec. 3.1, we evaluate our model across four dimensions (see the **supplementary material** for Gemini scoring details):

(1) Background Appearance Consistency, evaluated by the average per-frame **Mean Squared Error (MSE)**, **Structural Similarity (SSIM)** [43], and **Learned Perceptual Image Patch Similarity (LPIPS)** [52] between ground truth and output videos to assess restored scene fidelity;

(2) Motion Consistency (**Motion Cons.**), evaluated via CoTracker3 [18] through motion trajectory comparison between ground truth and output videos for the Cine-Restore subset, and via Gemini-based perceptual scoring for the Cine-NBG subset, ensuring background dynamics strictly follow foreground camera motion;

(3) Foreground Fidelity and Relighting, evaluated via Gemini-based perceptual scoring, using **Foreground Preservation (Fg. Pres.)** to quantify foreground identity preservation, and **Illumination Harmony (Illum. Harmony)** to assess the effectiveness of scene-adaptive relighting;

(4) Seamless Integration, also evaluated via Gemini-based perceptual scoring, using **Edge Quality** to assess the model’s ability to handle boundary imperfections, and **Foreground-Background Fusion (FG-BG Fusion)** to quantify

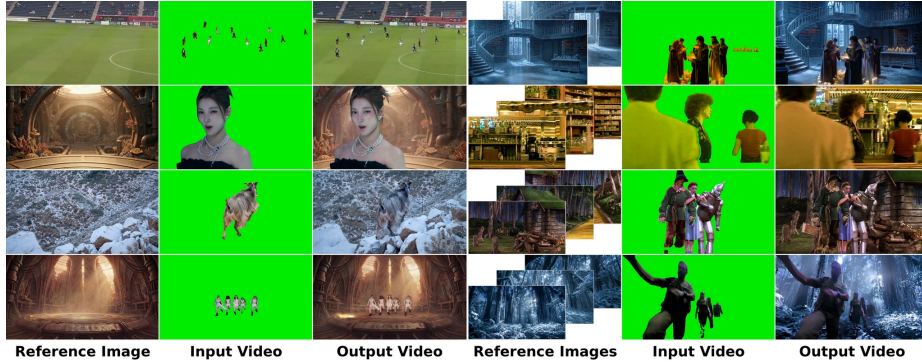


Fig. 4: Generation results of PAI-Studio. Readers can click and play the video clips in this figure using [Adobe Acrobat](#).

the global geometric, photometric, and temporal coherence at the foreground–background interface, ensuring a visually seamless composition.

4.2 Results

The generation results are shown in Figure 4. A key advantage of our approach is its strong temporal stability across long sequences. Unlike existing editing-based methods that suffer from flickering or static background artifacts, our model produces camera-aware background motion that closely follows foreground cinematographic dynamics, significantly reducing compositing artifacts even under minor matting imperfections. Furthermore, our framework achieves scene-consistent foreground relighting. When background references exhibit distinct lighting conditions—such as strong directional illumination or low-light environments—the synthesized videos maintain globally coherent shading and color adaptation on foreground subjects, preventing visual detachment between layers. Crucially, despite the global generative formulation, foreground identity and fine-grained textures are rigorously preserved across frames, avoiding the structural distortions and identity drift commonly observed in diffusion-based baselines. Finally, multi-reference background conditioning proves essential in large-motion scenarios. By aligning references from different temporal locations via positional encoding cloning, the model maintains appearance consistency under substantial viewpoint changes. Overall, explicitly modeling camera-aware motion and scene-consistent illumination enables high-quality cinematic background synthesis beyond existing approaches. More video results on edge quality and robustness to imperfect foregrounds are in the **supplementary material**.

4.3 Comparison and Evaluation

Quantitative Evaluation. Tables 2 and 3 report quantitative comparisons on the Cine-Restore and Cine-NBG subsets. Our method consistently outperforms both

Table 2: Quantitative comparison on the Cine-Restore subset under both single-frame and multi-frame settings. The last three metrics are evaluated via Gemini-based perceptual scoring (**scale 1-10**), whereas the rest are standard algorithmic metrics. Multi-frame results are reported only for methods that support multi-frame control.

Method	MSE ↓		SSIM ↑		LPIPS ↓		Motion Cons. ↑		Edge Quality ↑		FG-BG Fusion ↑		Illum. Harmony ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
VACE (Open-source)	0.0270	—	0.595	—	0.393	—	0.792	—	7.19	—	7.33	—	7.67	—
Kling O1 (Commercial)	0.0282	0.0190	0.646	0.682	0.402	0.341	0.706	0.721	8.44	9.04	7.88	8.20	7.62	7.44
Kling 3.0 Omni (Commercial)	0.0288	0.0210	0.592	0.622	0.414	0.370	0.787	0.715	8.12	8.40	7.68	7.96	7.92	7.44
Runway Aleph (Commercial)	0.0692	—	0.413	—	0.657	—	0.323	—	5.54	—	4.38	—	5.88	—
Switchx (Commercial)	0.0320	—	0.668	—	0.415	—	0.676	—	7.08	—	6.36	—	7.00	—
Ours	0.0164	0.00645	0.717	0.789	0.303	0.197	0.872	0.894	9.17	9.62	8.46	9.33	9.11	9.61

Table 3: Gemini-based perceptual evaluation on the Cine-NBG subset. Only Gemini-based scores (**scale 1-10**) are reported as no ground-truth video is available.

Method	Motion Cons. ↑	Edge Quality ↑	Fg. Pres. ↑	FG-BG Fusion ↑	Illum. Harmony ↑
VACE (Open-source)	6.92	7.82	2.44	7.04	8.38
Kling O1 (Commercial)	7.12	8.04	6.48	7.36	7.72
Runway Aleph (Commercial)	6.21	7.21	6.38	6.29	7.29
Kling 3.0 Omni (Commercial)	7.32	7.80	5.96	6.76	8.52
Switchx (Commercial)	7.12	7.40	8.88	6.16	7.88
Ours	8.16	8.86	9.58	8.40	9.22

open-source baselines and commercial APIs across all evaluation metrics. Specifically, our leading scores in Motion Consistency, alongside reconstruction metrics (MSE, SSIM, and LPIPS), demonstrate that our model generates camera-aware background dynamics while maintaining strict appearance consistency with the reference scene. Notably, multi-frame conditioning via positional encoding cloning (Table 2) further improves these results, validating its effectiveness for handling large camera motions. Furthermore, our approach effectively addresses scene-adaptive relighting and imperfect foreground boundaries, as evidenced by top performance in Foreground Preservation (Fg. Pres.), Illumination Harmony, Edge Quality, and FG-BG Fusion. These results confirm that our model can naturally integrate actors into the target scene for both existing background reconstruction (Table 2) and new background replacement tasks (Table 3).

Qualitative Evaluation. Figure 5 presents qualitative comparisons. Existing video editing methods often generate static or weakly dynamic backgrounds, leading to noticeable inconsistencies when the camera moves. Baseline results typically exhibit misalignment artifacts, inconsistent lighting and noticeable foreground identity distortion, as highlighted in the **red boxes** in Figure 5. In contrast, our method produces backgrounds that evolve coherently over time and exhibit illumination consistent with the target scene. Crucially, our approach maintains strict foreground consistency, faithfully preserving the actors’ identities and structural details without degradation. The foreground actors appear naturally integrated into the synthesized environment, with reduced compositing artifacts and strong robustness across various scenarios.



Fig. 5: Comparison results show that our method consistently outperforms baseline approaches in structural fidelity, illumination harmony, and temporal coherence. Readers can click and play the video clips in this figure using [Adobe Acrobat](#).

Table 4: Ablation study on the Cine-Restore subset. We evaluate each component under both single-frame and multi-frame control protocols.

Method	MSE ↓		SSIM ↑		LPIPS ↓		Motion Cons. ↑		Edge Quality ↑		FG-BG Fusion ↑		Illum. Harmony ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
w/o Position Encoding Control	0.0300	0.0160	0.608	0.681	0.399	0.287	0.824	0.843	6.67	8.38	6.85	8.41	7.63	8.88
w/o Illumination Adaptation	0.0270	0.0130	0.607	0.681	0.381	0.279	0.850	0.879	6.40	7.25	6.39	7.24	6.44	7.31
w/o Structured Annotation	0.0177	0.00695	0.715	0.787	0.307	0.198	0.873	0.902	9.10	9.52	8.39	9.21	8.88	9.57
Full Model	0.0164	0.00645	0.717	0.789	0.303	0.197	0.872	0.894	9.17	9.62	8.46	9.33	9.11	9.61

4.4 Ablation Study

We conduct ablation experiments on the Cine-Restore subset to verify our core designs. Qualitative results are illustrated in Fig. 6. As shown in Table 4, the Full Model achieves the best or near-best performance across all metrics. Without Illumination Adaptation, the model fails to automatically adjust foreground lighting to match the target environment, resulting in poor foreground-background integration and notably lower scores in illumination harmony, visually evident from the foreground lighting mismatch with the ground truth in Fig. 6(a). Without Position Encoding Control, although the model produces normal-looking frames, it loses the ability to map background references to precise temporal locations, leading to a significant drop in motion consistency and temporally controllable synthesis, visually evident from the inaccurate camera tilt and road shifting in the top example of Fig. 6(b). Furthermore, relying on simple text prompts rather than our structured annotations degrades quantitative performance: the model misses critical illumination and semantic cues, manifesting as mismatched lighting or background collapse in Fig. 6(c). Notably, the richer fine-grained captions in our annotations improve robustness even with incom-

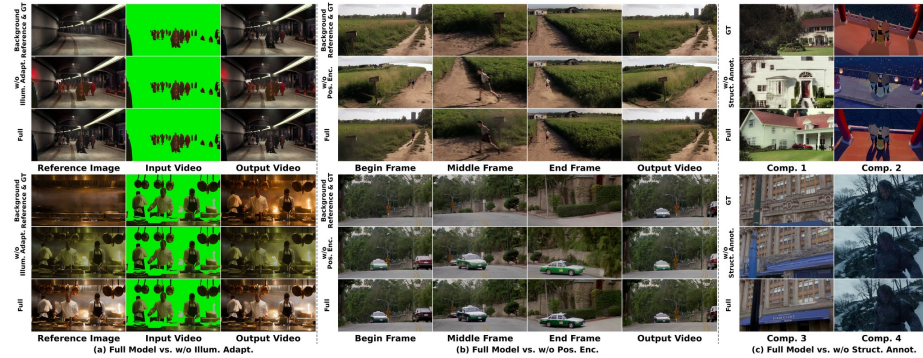


Fig. 6: Ablation study results. The full model achieves the best or near-best performance, validating the importance of position encoding for motion consistency, illumination adaptation for realistic integration, and structured annotation for detailed guidance. Readers can click and play the videos in this figure using [Adobe Acrobat](#).

plete keyframe cues, yielding results closer to the ground truth. These results confirm that precise temporal alignment, active relighting, and detailed textual guidance are essential for realistic video synthesis.

4.5 User Study

We conduct a user study to evaluate perceptual quality and realism from a human perspective, which is particularly important for virtual studio and film production scenarios. We randomly sample 50 videos from both the Cine-Restore and Cine-NBG subsets, and compare our method with representative open-source methods and commercial APIs. A total of 25 participants with prior experience in video editing or visual content creation took part in the study. Each test case is evaluated by multiple independent participants, and the final scores are reported as the percentage of times each method is voted as the best result. As shown in Figure 7, our method is consistently preferred across all five evaluation criteria, aligning with our quantitative evaluations and confirming that our model comprehensively better matches human judgments across all dimensions of the cinematic background replacement task. Our method receives the largest vote share on all five criteria (29.7%–35.6%), clearly ahead of the strongest baseline, Kling O1 (around 21%–24%). The display order is randomized in every trial (a chi-square test confirms no position bias), and pairwise significance tests confirm that our lead is statistically significant ($p < 0.01$) across all dimensions; full per-pair win rates are reported in the supplementary material.

5 Conclusion

We introduced PAI-Studio, a multi-condition video synthesis task for cinematic background replacement that jointly requires motion-aligned background genera-

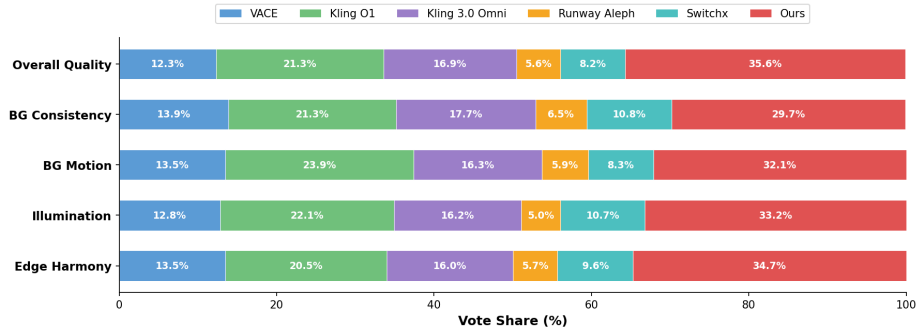


Fig. 7: User study results comparing our method with open-source methods and commercial APIs. Values indicate the percentage of votes received by each method.

tion, scene-consistent foreground relighting, and robustness to imperfect segmentation boundaries. We presented an in-context Diffusion Transformer framework that concatenates foreground video, multi-frame background references, and denoising tokens into a unified sequence and models them with full bidirectional attention for coherent cross-condition reasoning. Temporal positional encoding cloning enables precise multi-frame temporal control, while foreground relighting and color perturbations during training encourage illumination recovery from scene cues. We further curated a 30K-scale dataset from films and online videos with structured motion annotations. Extensive quantitative and qualitative results demonstrate significant improvements over open-source and commercial API baselines, and user studies confirm a strong preference for our outputs. We believe this work provides a practical foundation toward film-quality controllable video generation and virtual production.

Acknowledgements

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023), and by the College of Design and Engineering at the National University of Singapore.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bar-Tal, O., Ofri-Amar, D., Fridman, R., Kasten, Y., Dekel, T.: Text2live: Text-driven layered image and video editing. In: European conference on computer vision. pp. 707–723. Springer (2022)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion mod-

- els. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
4. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. *OpenAI Blog* 1(8), 1 (2024)
5. Cai, M., Cun, X., Li, X., Liu, W., Zhang, Z., Zhang, Y., Shan, Y., Yue, X.: Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7763–7772 (2025)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025)
7. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512* (2023)
8. Chen, P., Zeng, X., Zhao, M., Ye, P., Shen, M., Cheng, W., Yu, G., Chen, T.: Sparse-vdit: Unleashing the power of sparse attention to accelerate video diffusion transformers. *arXiv preprint arXiv:2506.03065* (2025)
9. Chen, Y., Zhang, J., Hu, T., Zeng, Y., Xue, Z., He, Q., Wang, C., Liu, Y., Hu, X., Yan, S.: Ivebench: Modern benchmark suite for instruction-guided video editing assessment. *arXiv preprint arXiv:2510.11647* (2025)
10. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7346–7356 (2023)
11. Gao, P., Zhuo, L., Liu, D., Du, R., Luo, X., Qiu, L., Zhang, Y., Lin, C., Huang, R., Geng, S., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. *arXiv preprint arXiv:2405.05945* (2024)
12. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373* (2023)
13. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022)
14. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. *Iclr* 1(2), 3 (2022)
16. Hu, Z., Xu, D.: Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073* (2023)
17. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
18. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6013–6022 (2025)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024)

20. Li, Z., Lu, C.Z., Qin, J., Guo, C.L., Cheng, M.M.: Towards an end-to-end framework for flow-guided video inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17562–17571 (2022)
21. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021)
22. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 238–247 (2022)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4d creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025)
25. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
26. Ma, Y., He, Y., Wang, H., Wang, A., Shen, L., Qi, C., Ying, J., Cai, C., Li, Z., Shum, H.Y., et al.: Follow-your-click: Open-domain regional image animation via motion prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6018–6026 (2025)
27. Ma, Y., Liu, H., Wang, H., Pan, H., He, Y., Yuan, J., Zeng, A., Cai, C., Shum, H.Y., Liu, W., et al.: Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–12 (2024)
28. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. arXiv preprint arXiv:2506.05207 (2025)
29. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
30. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
31. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023)
32. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: International Conference on Learning Representations (2023)
33. Singer, U., Zohar, A., Kirstain, Y., Sheynin, S., Polyak, A., Parikh, D., Taigman, Y.: Video editing via factorized diffusion distillation. In: European Conference on Computer Vision. pp. 450–466. Springer (2024)
34. Song, Q., Shen, Y., Chen, M., Sun, H., Lan, J., Zhu, X., Zheng, B., Cao, L.: Fashionchameleon: Towards real-time and interactive human-garment video customization. arXiv preprint arXiv:2605.15824 (2026)
35. Song, Y., Huang, S., Yao, C., Ye, X., Ci, H., Liu, J., Zhang, Y., Shou, M.Z.: Processpainter: Learn painting process from sequence data. arXiv preprint arXiv:2406.06062 (2024)
36. Song, Y., Liu, C., Jiang, Y., Shou, M.Z.: Streamingeffect: Real-time human-centric video effect generation. arXiv preprint arXiv:2605.17019 (2026)

37. Song, Y., Liu, C., Shou, M.Z.: Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. *arXiv preprint arXiv:2502.01572* (2025)
38. Song, Y., Yao, W., Wang, H., Shou, M.Z.: Vista: Triplet-supervised video style transfer with diffusion transformers. *arXiv preprint arXiv:2605.17312* (2026)
39. Sun, Y., Wang, G., Gu, Q., Tang, C.K., Tai, Y.W.: Deep video matting via spatio-temporal alignment and aggregation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6975–6984 (2021)
40. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023)
41. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
42. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: *Advances in Neural Information Processing Systems* (2018)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
44. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 7623–7633 (2023)
45. Xing, J., Mai, L., Ham, C., Huang, J., Mahapatra, A., Fu, C.W., Wong, T.T., Liu, F.: Motioncanvas: Cinematic shot design with controllable image-to-video generation. In: *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. pp. 1–11 (2025)
46. Xing, J., Xia, M., Zhang, Y., Chen, H., Yu, W., Liu, H., Liu, G., Wang, X., Shan, Y., Wong, T.T.: Dynamicrafter: Animating open-domain images with video diffusion priors. In: *European Conference on Computer Vision*. pp. 399–417. Springer (2024)
47. Yan, H., Song, X., Wang, Y., Li, H., Ji, P., Ma, C.: Bachvid: Training-free video generation with consistent background and character. *arXiv preprint arXiv:2510.21696* (2025)
48. Yan, W., Brown, A., Abbeel, P., Girdhar, R., Azadi, S.: Motion-conditioned image animation for video editing. *arXiv preprint arXiv:2311.18827* (2023)
49. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Rerender a video: Zero-shot text-guided video-to-video translation. In: *SIGGRAPH Asia 2023 Conference Papers*. pp. 1–11 (2023)
50. 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)
51. Zhang, L., Agrawala, M.: Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113* (2024)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 586–595 (2018)
53. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 10477–10486 (2023)