FILMTOGETHER: CINEMATIC VIDEO GENERATION FROM LOW-COST INPUTS

Shiying Zhang **Xiaobin Hu** **Qianchi Zhang** **Jiaqi Zhao**
Xiaoda Yang Bo Yin Shuicheng Yan
National University of Singapore

ABSTRACT

The widespread adoption of mobile devices has democratized video recording, but a stark aesthetic divide persists between casual, low-cost captures and high-cost cinematic productions. Bridging this gap requires upgrading heterogeneous production elements, such as lighting, specialized costumes, and materials, without altering the core layout, identity, and temporal motion of the original footage. However, this task is heavily bottlenecked by two important challenges: 1) Lacking a high-quality paired training data specifically designed to handle complex low-to-high appearance shifts. 2) Ignoring the capacity of current video editing models to perceive and enhance the multiple low-cost compositions. To address these issues, we present CineCraft-120K, a large-scale dataset of aligned low- and high-cost video pairs annotated with pair-aware editing prompts. CineCraft-120K features a diverse, high-resolution and high-cost **Cine**matic **Craft**ing bank spanning decades of media production, meticulously taxonomized into **5** broad domains and **23** fine-grained sub-categories. It also provides element-level mapping between low- and high-cost counterparts, offering pair-aware rich textual prompts that precisely detail the required aesthetic transformations while anchoring structural layout. Architecturally, we propose a simple yet effective dual-memory feature adapter, where a text-memory expands semantic capacity for low2high appearance changes, and a hint memory supplies layer-wise visual feature corrections of high-cost element before context injection. Extensive experiments and ablation studies demonstrate the superiority of CineCraft-120K for low-to-high-cost video editing, alongside the effectiveness of our dual-memory adapter.

Code: `GitHub` **Data:** `CineCraft-120K` (Hugging Face)

1 INTRODUCTION

The rapid evolution of mobile technology has made video recording accessible to everyone, fundamentally changing how everyday moments are captured and shared. However, a persistent visual gap remains between these casual, low-cost videos and professional cinematic productions. Low-cost videos are typically characterized by amateur framing, ordinary settings, and suboptimal lighting, whereas high-cost productions benefit from professional equipment, sophisticated illumination, and intricate set designs. Advances in denoising, latent, video, and transformer-based diffusion models (Ho et al., 2020; Song et al., 2020a;b; Rombach et al., 2022; Ho et al., 2022; Peebles & Xie, 2023) have propelled instruction-based video editing to new heights (Cheng et al., 2024; Wu et al., 2025b; Bai et al., 2026), yet transforming a basic low-cost video into a cinematic masterpiece remains a largely unexplored frontier. Bridging this aesthetic divide is not a simple global style transfer, but it requires upgrading heterogeneous production elements—such as lighting, specialized costumes, and materials—without altering the core layout, identity, and temporal motion of the original footage.

Realizing this low-to-high cost transformation is heavily bottlenecked by two primary challenges: *1)* Lacking high-quality paired training data for complex appearance shifts. The progress of specialized cinematic editing is hampered by the scarcity of appropriate training data. Current video editing datasets often emphasize generic instructions or broad stylization (Zi et al., 2026; He et al., 2025; Huang et al., 2026). They lack the domain-specific mappings and task diversity required to supervise the intricate transition from low-cost to high-cost aesthetics while strictly preserving

spatial layout and motion. There is a distinct absence of large-scale, paired data that explicitly captures these element-level production upgrades. *2)* Ignoring the capacity of current models to better perceive and enhance multiple low-cost compositions. At the model level, existing architectures struggle to strengthen the heterogeneous and multiple low-cost compositions. Standard generators (*e.g.,* Ditto (Bai et al., 2026)) typically inject source visual hints directly into a frozen denoising backbone, leaving the model with limited capacity to process semantic instructions related to complex appearance changes, and failing to reliably disentangle the core structure from the low-cost aesthetic.

In this work, we present *CineCraft-120K*, a carefully curated, high-quality dataset comprising over 120K premium cinematic video clips, with over 99% of the clips feature a resolution of at least 1080*p* and are accompanied by detailed captions. As shown in Fig. 1 and 2, our CineCraft-120K has several characteristics: *1). Superior visual diversity and temporal coverage*: The cinematic crafting bank distinguishes itself through exceptional visual diversity, meticulously taxonomized into 5 broad domains and 23 fine-grained sub-categories, while encompassing over five decades of media production from 1970 to 2023. This extensive temporal coverage prevents the model from overfitting to contemporary trends, ensuring it learns generalized and enduring cinematic aesthetics. *2). High-quality video pairs via a scalable dataset pipeline*: Supported by approximately 1,500 GPU days computational source, it features over 120K high-resolution, aligned low- and high-cost video pairs generated through a highly efficient, automated construction pipeline. Through strict multi-dimensional quality filtering, this intensive approach ensures superior visual richness and clarity, with over 99% of the resulting cinematic clips featuring 1080p or 4K resolutions. *3). Expressive in pair-aware rich textual captions*: Unlike datasets that provide generic video descriptions, it introduces highly detailed, pair-aware element-wise prompts averaging around 125 dense words. These expressive instructions are grounded in the actual visual transformations between counterparts, precisely detailing coordinated upgrades to heterogeneous elements like lighting, color grading, and materials.

To address the architecture deficit, we propose a dual-memory adapter that seamlessly maps low-cost inputs to cinematic quality with minimal overhead. At the semantic level, the Instruction-Routed Text Memory expands the instruction representation capacity via learnable text-memory token, overcoming the limitations of global textual prompts in detailing diverse production elements. In parallel, the Layer-Wise Hint Memory operates at the visual context injection to supply layer-specific corrections to the source video hints before DiT fusion. This adaptive feature refinement prevents the generative process from being bottlenecked by the coarse textures and defects of the source video, enabling multi-level enhancements from low-level aesthetics to high-level layouts. Synergistically, the text memory deeply enriches the cinematic semantics while the hint memory finely calibrates the visual context, facilitating sophisticated element-level upgrades with high efficiency.

Our contributions are listed as follows: 1). We introduce *CineCraft-120K*, a large-scale, high-quality dataset comprising over 120K aligned low- and high-cost video pairs with pair-aware prompts for cinematic video editing. 2). We propose a Dual-Memory Adapter, which combines text and layer-wise hint memories to seamlessly execute complex aesthetic transformations with minimal overhead. 3). We design a VLM-Aesthetic evaluator and demonstrate the superiority of our framework on the low-to-high-cost video editing task.

2 CINECRAFT-120K: PAIRED CINEMATIC DATA CONSTRUCTION

We introduce CineCraft-120K, a large-scale paired dataset for upgrading low-cost videos into cinematic productions while preserving their underlying scene structure and motion. Its construction follows three design principles: diverse cinematic scenarios, structure-preserving counterpart synthesis, and pair-aware, element-level instruction caption annotations. Starting from a curated high-cost visual bank, we synthesize low-cost counterparts and subsequently annotate the reverse transformation through precise comparison of each video pair. This approach retains real cinematic footage as the enhancement target while providing explicit supervision for coordinated changes to heterogeneous production elements. Fig. 1 illustrates the complete pipeline.

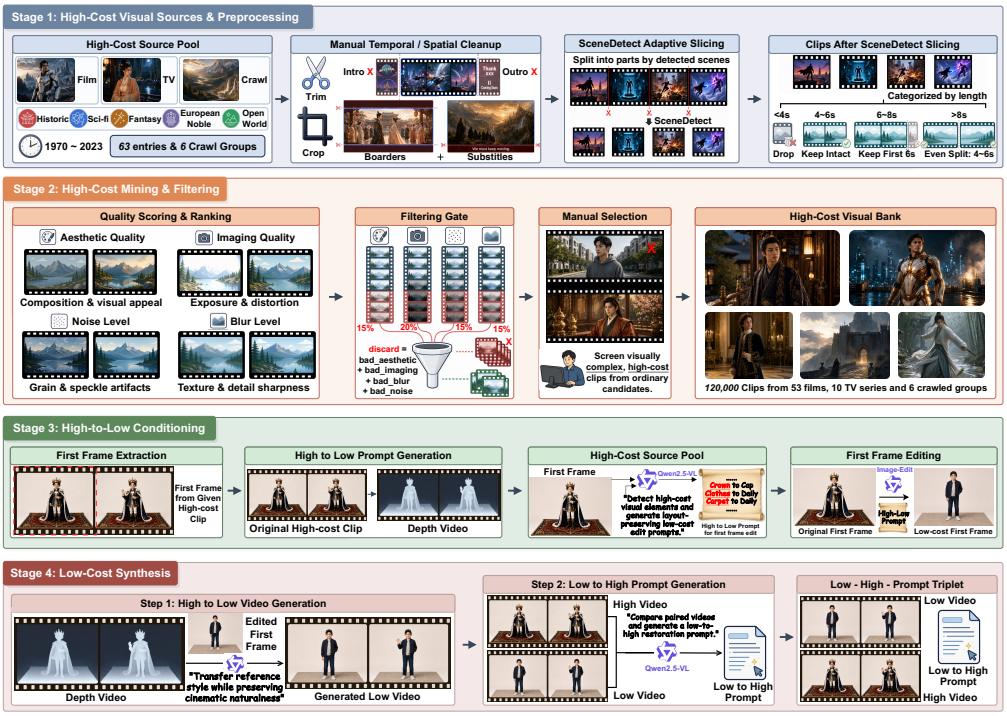


Figure 1: CineCraft-120K construction pipeline. We curate a high-cost visual bank, synthesize aligned low-cost counterparts through first-frame editing and depth-guided generation, and verify and adapt editing prompts to describe the observed low-to-high transformations.

2.1 Cinematic Videos Curation and Pair-Aware Data Construction

Diverse Cinematic Source Collection. To capture a broad range of cinematic appearances, we collect visually rich footage from films, television series, and crawled open-world videos. The source collection spans media released between 1970 and 2023 and is organized into five broad visual domains and 23 fine-grained categories. This coverage exposes the dataset to varied production designs, lighting styles, costumes, and materials across genres and eras, broadening its scope beyond contemporary cinematic appearances. The crawled subset further extends the collection to visual content beyond narrative film and television.

Shot-Coherent Preprocessing. We first remove openings, endings, black bars, subtitle regions, and other non-content areas using manually specified temporal boundaries and spatial crops. We then apply adaptive shot detection with PySceneDetect(Castellano, n.d.) to obtain temporally coherent clips without crossing detected shot boundaries. Clips shorter than 4 seconds are discarded; those between 4 and 6 seconds are retained; those longer than 6 but shorter than 8 seconds are truncated to their first 6 seconds; and those of at least 8 seconds are evenly divided into approximately 4–6-second segments. This procedure produces compact source clips while avoiding short residual segments and preserving within-shot continuity.

Multi-Dimensional Quality Curation. We combine manual screening for visually rich, high-production-value content with automated perceptual quality filtering. VBench (Huang et al., 2024) evaluates aesthetic and imaging quality, for which we reject the bottom 15% and 20% of candidates, respectively. FineVQ (Duan et al., 2025) provides complementary blur- and noise-related quality assessments, with the lowest-quality 15% rejected under each criterion. A clip is removed if it falls into the rejection set of any criterion. This filtering strategy screens for complementary failure modes, including weak composition, exposure and distortion artifacts, excessive noise, and blurred details. The retained clips form the high-cost visual bank that supplies cinematic targets for subsequent pair construction.

Structure-Guided High-to-Low Synthesis. To construct paired supervision, we generate a low-cost counterpart for each selected high-cost clip by simplifying its production-related appearance under structural guidance. From each reference clip, we extract the first frame and estimate a depth

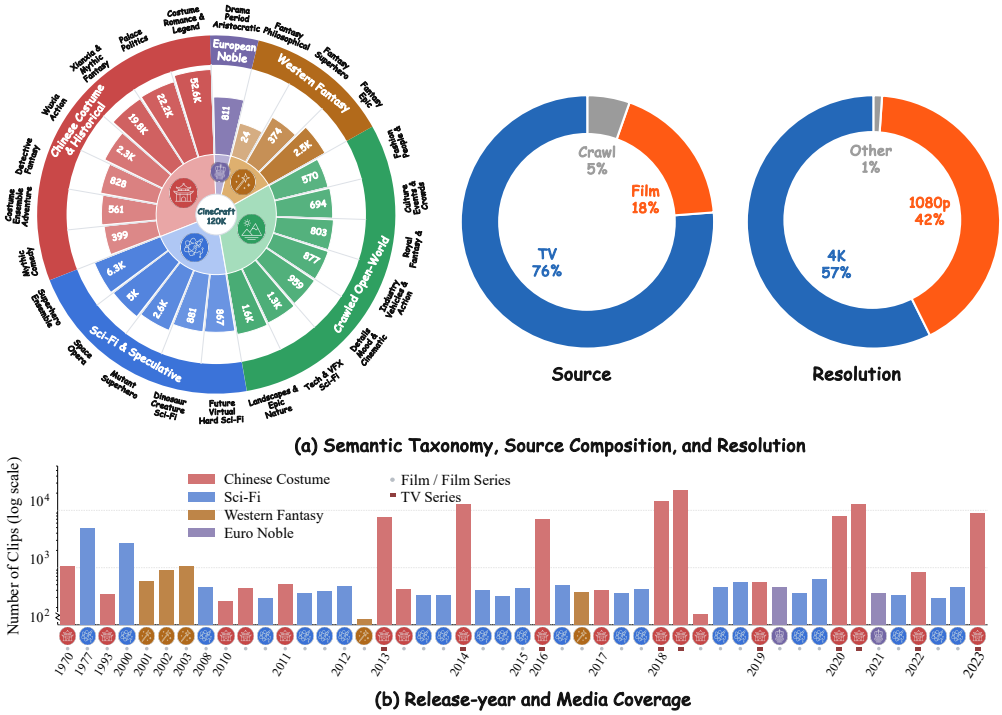


Figure 2: Source composition and visual coverage of CineCraft-120K.

sequence using Video-Depth-Anything (Chen et al., 2025). Qwen2.5-VL-32B-Instruct (Bai et al., 2025b) examines the first frame and generates a high-to-low editing instruction that identifies visually elaborate elements and specifies replacements while retaining their spatial arrangement. Qwen-Image-Edit (Wu et al., 2025a), implemented through DiffSynth-Studio, applies this instruction to produce a low-cost first frame. Wan2.1-VACE-14B (Jiang et al., 2025; Wan et al., 2025) then synthesizes the low-cost video conditioned on the edited frame and reference depth sequence. The edited frame establishes the intended low-cost appearance, while the depth sequence guides scene geometry and temporal motion. The overall construction effort uses approximately 1,500 GPU-days. This process supports large-scale pair generation with appearance changes guided by a shared underlying scene structure.

Prompt-Guided Verification and Adaptive Annotation. We use the preceding high-to-low editing instruction as a reference to verify whether the synthesized low-cost video realizes the intended appearance changes. Specifically, Qwen3-VL-8B-Instruct jointly examines the editing instruction, the original high-cost video, and its generated low-cost counterpart to assess whether the specified element-level modifications are reflected in the resulting pair. This comparison identifies discrepancies between the intended and realized edits, including changes that are missing, only partially achieved, or introduced beyond the instruction.

When discrepancies are detected, the instruction is adaptively revised to reflect the transformations actually observed in the video pair. The verified and adjusted transformation description is then expressed as a low-to-high enhancement instruction, specifying how to upgrade the generated low-cost video toward its high-cost reference. This verification-and-adaptation process grounds the final supervision in both the original editing intent and the realized visual changes, reducing mismatches between textual instructions and paired video content.

2.2 Dataset Characteristics: Diversity, Quality, and Expressiveness

We characterize CineCraft-120K through visual coverage, source quality, and pair-aware instructions. Its high-cost bank contains **124,902 clips** totaling **176.38 hours**; Figures 2 and 3 summarize its statistics.

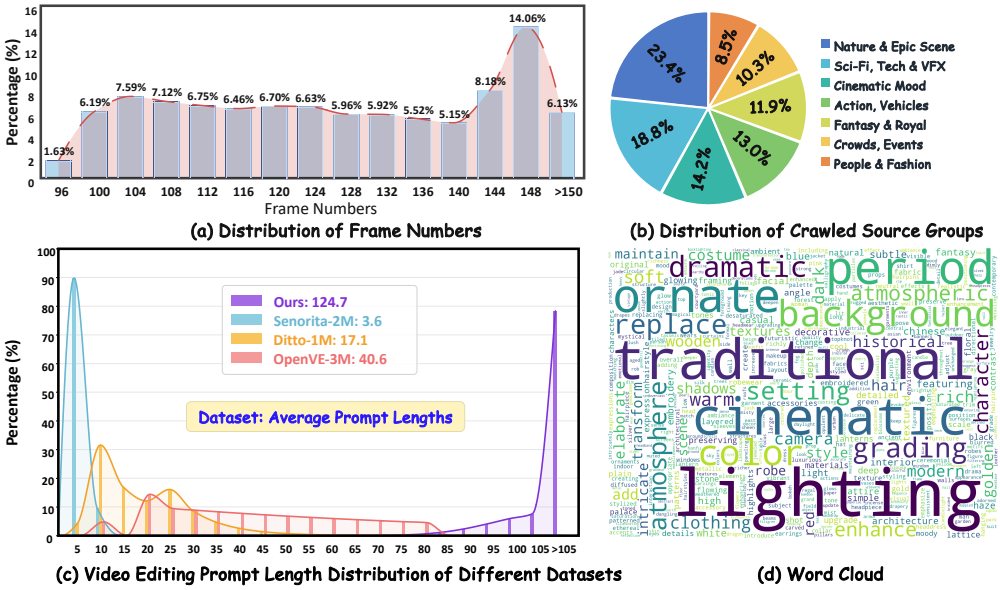


Figure 3: CineCraft-120K statistics: (a) source-clip frame-count distribution; (b) crawled-content categories; (c) editing-prompt length distributions; and (d) editing-instruction word cloud.

Visual Diversity and Historical Coverage. CineCraft-120K spans 5 broad domains and 23 fine-grained categories. The taxonomy covers Chinese costume and historical scenes (7 categories), sci-fi and speculative content (5), Western fantasy (3), European aristocratic scenes, and crawled open-world content (7), from wuxia and palace settings to space opera and epic landscapes (Fig. 2(a)).

Sources span 1970–2023, while crawled clips extend beyond film and television. Their largest visual groups are nature/epic scenes (23.4%), sci-fi/technology/VFX (18.8%), cinematic mood (14.2%), and action/vehicles (13.0%); the remaining groups are fantasy/royalty (11.9%), crowds/events (10.3%), and people/fashion (8.5%) (Figs. 2(b) and 3(b)).

Large-Scale, High-Quality Cinematic Paired Videos. Television contributes about 76% of the visual bank, and 99% of source clips are 1080p or 4K. Combined with multi-dimensional quality filtering, these sources provide detailed appearance references; Fig. 3(a) shows their frame-count distribution.

Expressive, Pair-Aware Editing Captions. Prompt-guided verification grounds each low-to-high instruction in the edits realized by its video pair. Instructions average **124.7 words** (median **122**; IQR **108–138**) and describe coordinated changes to lighting, color, clothing, materials, and scene appearance. Their length and vocabulary distributions appear in Figs. 3(c,d).

3 Methodology: Dual-Memory Adaptation

Given a low-cost source video and an editing instruction, FilmTogether aims to enhance cinematic appearance while preserving source identity, spatial layout, and temporal motion. This task requires coordinated changes to heterogeneous production elements, including illumination, costumes, materials, and set design. It therefore places complementary demands on the conditioning pathways: textual features must represent the intended appearance changes, while visual features must retain structural guidance without restricting generation to the source appearance.

FilmTogether builds on Wan2.1-VACE-14B (Wan et al., 2025; Jiang et al., 2025).

As shown in Fig. 4, we introduce a **dual-memory adapter** that augments both conditioning pathways while keeping the pretrained backbone frozen.

Instruction-Routed Text Memory supplements the instruction with learnable, adaptively gated tokens.

Layer-Wise Hint Memory uses the augmented instruction to retrieve layer-specific residuals that refine source-video hints before their injection into the transformer-based DiT Main Branch (Vaswani et al., 2017; Peebles & Xie, 2023).

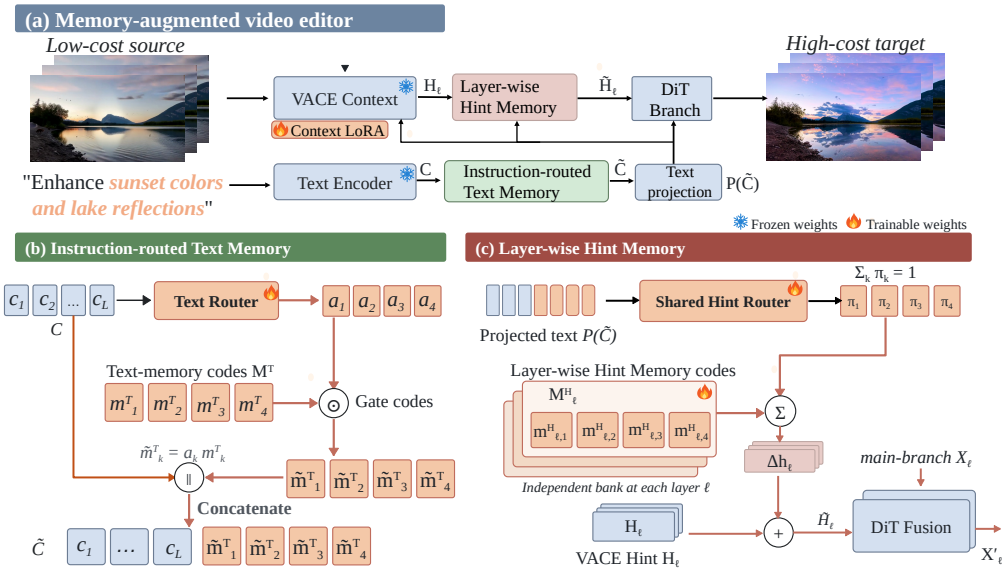


Figure 4: FilmTogether architecture: (a) overall framework; (b) instruction-routed Text Memory enriches semantic conditioning; and (c) layer-wise Hint Memory refines source visual hints before DiT fusion.

Instruction-Routed Text Memory. The frozen text encoder provides general-purpose instruction features but cannot directly adapt its parameters to cinematic enhancement. We introduce a learnable memory that expands the conditioning representation while retaining the original instruction tokens.

For clarity, let $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_L] \in \mathbb{R}^{B\times L\times d_T}$ denote the frozen text-encoder features and $\bar{\mathbf{c}} = L^{-1}\sum_i \mathbf{c}_i$ their pooled representation. The learnable text-memory codes are $\mathbf{M}^T = \{\mathbf{m}_k^T\}_{k=1}^K$ ($K = 4$), each with the same dimension as a text token. A lightweight router predicts independent sigmoid gates:

$$\mathbf{a}^T = \mathrm{sigmoid}\big(\mathbf{W}_2^T\,\phi(\mathbf{W}_1^T\,\mathrm{LN}(\bar{\mathbf{c}}))\big)\,, \qquad 0 \le a_k^T \le 1. \tag{1}$$

where $\bar{\mathbf{c}}$ is the mean-pooled instruction feature, LN is layer normalization (Ba et al., 2016), ϕ is GELU (Hendrycks & Gimpel, 2016), sigmoid independently maps each router logit to $[0, 1]$, $\mathbf{W}_1^T$ and $\mathbf{W}_2^T$ are the trainable Text-Router projections, k indexes a memory code, and the superscript T denotes Text Memory. The gates scale the codes before appending them to the original text sequence:

$$\widetilde{\mathbf{C}} = [\mathbf{C}; a_1^T\mathbf{m}_1^T; \ldots; a_K^T\mathbf{m}_K^T]. \tag{2}$$

where $\widetilde{\mathbf{C}} \in \mathbb{R}^{B\times(L+K)\times d_T}$ is the augmented instruction sequence, B is the batch size, L is the instruction length, d_T is the text-feature dimension, $\mathbf{c}_i \in \mathbb{R}^{d_T}$ is the i-th frozen text token, $K = 4$ is the number of learnable codes, $\mathbf{m}_k^T \in \mathbb{R}^{d_T}$ is the k-th code, $[\,;\,]$ denotes token-wise concatenation.

Independently gated memory tokens expand semantic capacity for coordinated appearance upgrades while retaining the original instruction tokens. Unlike softmax selection, the independent gates can activate several codes simultaneously for instructions that combine lighting, material, costume, and set-design changes, while appending the codes preserves every literal instruction token.

Layer-Wise Hint Memory. The VACE Context Branch extracts source-video hints $\{H_\ell\}$ at designated fusion locations. These features carry structural information useful for preserving the source, but may also encode its low-cost textures, lighting, and other appearance characteristics. Hint Memory introduces an instruction-conditioned residual at each fusion location, allowing the visual conditioning to adapt before entering the main branch.

$$\boldsymbol{\pi}^H = \mathrm{softmax}\Big(\mathbf{W}_2^H\,\phi\Big(\mathbf{W}_1^H\,\mathrm{LN}(\mathrm{Pool}(P(\widetilde{\mathbf{C}})))\Big)\Big)\,, \qquad \sum_{k=1}^{K}\pi_k^H = 1. \tag{3}$$

where P is the frozen projection from text width d_T to the DiT hidden width, Pool averages the projected tokens, $\mathbf{W}_1^H$ and $\mathbf{W}_2^H$ are the trainable Hint-Router projections, $\boldsymbol{\pi}^H = [\pi_1^H, \ldots, \pi_K^H]$ is

the resulting instruction-conditioned distribution, and the superscript H denotes Hint Memory. Each fusion location ℓ owns an independent bank $\mathbf{M}_\ell^H = \{\mathbf{m}_{\ell,k}^H\}_{k=1}^K$ in the DiT hidden dimension. The shared routing decision reads a layer-specific correction:

$$\Delta\mathbf{h}_\ell = \sum_{k=1}^{K} \pi_k^H \mathbf{m}_{\ell,k}^H. \tag{4}$$

where ℓ indexes a DiT fusion location, $\mathbf{m}_{\ell,k}^H$ is the k-th code in that bank, π_k^H is its shared instruction-dependent weight, and $\Delta\mathbf{h}_\ell$ is the resulting layer-specific correction.

Here the correction is broadcast to all hint tokens. Independent token banks let early, middle, and late fusion locations learn different corrections. Its interaction with position-dependent features and subsequent attention yields non-uniform spatiotemporal effects. Placing the correction before fusion is important, after considering the fact that the original hint continues to anchor identity, geometry, and motion, while the routed residual removes the need for the main branch to inherit its low-cost appearance unchanged. Layer-specific token banks further allow early injections to refine local texture and illumination and later injections to adjust higher-level scene organization, without unfreezing the visual backbone.

Distilled Aesthetic Evaluator. Following knowledge distillation (Hinton et al., 2015), we distill Gemini 3.1 Pro Preview's aesthetic assessments of 5,000 edited clips into Qwen3-VL-8B-Thinking (Bai et al., 2025a) using LoRA (Hu et al., 2021) and token-level cross-entropy loss. Supervision includes five aspect scores, an overall score, and corresponding rationales. Used only for evaluation, the frozen model receives edited frames and weak instruction context. We report its overall score as VLM-Aesthetic; Appendix B.1 provides the full rubric.

4 Experiments and Results

4.1 Experimental Setup

Implementation details. Our model builds on Wan2.1-VACE-14B (Wan et al., 2025; Jiang et al., 2025) with Ditto's Context-Branch LoRA (Bai et al., 2026). We augment the pretrained editor with Text Memory and layer-wise Hint Memory, then fine-tune on the CineCraft-120K training split. The reference-frame prior is linearly annealed and disabled at inference. We use AdamW (Loshchilov & Hutter, 2017) with a learning rate of 5×10^{-5} and weight decay of 0.01. Both stages use bfloat16 and an effective batch size of eight.

Competitors. We compare *Filmtogether* against VACE (Jiang et al., 2025), Bernini-R-14B (Team et al., 2026), Ditto (Bai et al., 2026), InsV2V (Cheng et al., 2024), InsViE (Wu et al., 2025b), TokenFlow (Geyer et al., 2024), ICVE (Liao et al., 2025), Kiwi-Edit (Lin et al., 2026), SANA-Streaming (Zhao et al., 2026), Señorita (Zi et al., 2026), ReCo (Zhang et al., 2025), and Lucy-Edit (Team, 2025). For a fair comparison, we evaluate each baseline using the settings recommended in its paper or official implementation.

Evaluation metrics. We measure temporal consistency using CLIP-F (Radford et al., 2021) and DINO-F (Caron et al., 2021), which compute mean adjacent-frame feature cosine similarity, and text-conditioned preference using frame-averaged PickScore (Kirstain et al., 2023). For cinematic quality, we use the VLM-Aesthetic evaluator detailed in Appendix B.1 to deterministically score two edited frames sampled at 1 fps and resized to 640 pixels wide. We report its holistic score, a semantic-design score averaging narrative/emotional fit, style/world consistency, and composition/lighting, and separate color/texture and hierarchy/readability scores.

Cinematic Crafting Benchmark (CineCraft-Bench). CineCraft-Bench comprises 100 videos spanning Chinese costume and historical scenes, science fiction, Western fantasy, European aristocratic settings, and open-world content. These videos cover diverse subjects, environments, and cinematic styles, with editing instructions targeting lighting, color grading, costumes, materials, and scene appearance.

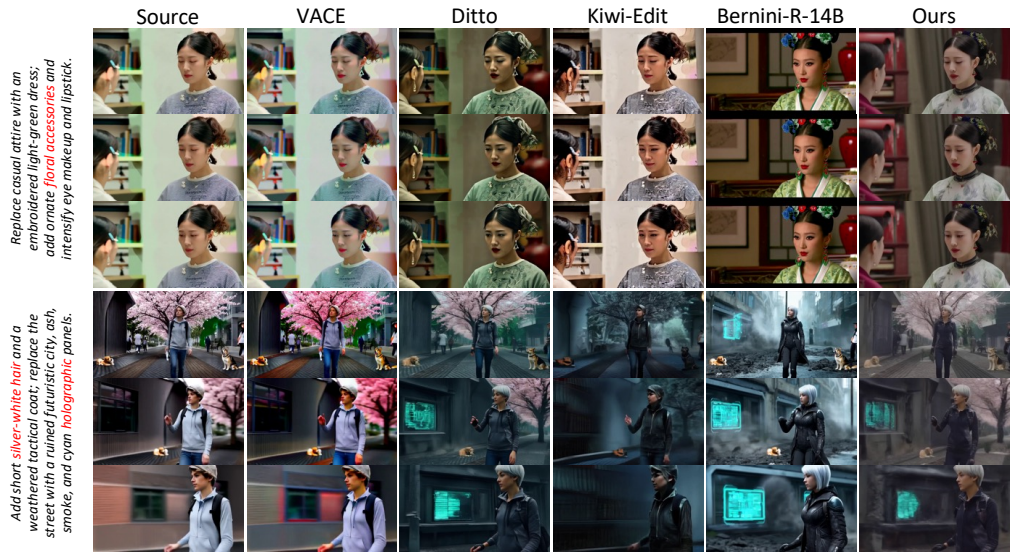


Figure 5: Qualitative comparison on two representative CineCraft-Bench cases. Each method receives the same source video and editing instruction, and three matched timestamps are shown. FilmTogether realizes the requested period costume and science-fiction transformations while preserving subject identity, pose, and temporal structure.

4.2 Quantitative Comparison

Table 1 compares FilmTogether with 12 state-of-the-art video editing models on low-to-high-cost cinematic enhancement. FilmTogether achieves the highest CLIP-F (0.9967) and DINO-F (0.9985), indicating strong temporal consistency during appearance transformation. Its VLM-Aesthetic score is 3.87, compared with 3.77 for Bernini-R-14B.

The fine-grained aesthetic scores further highlight its effectiveness for low-to-high enhancement. FilmTogether leads in narrative/style/lighting (3.88), color/texture (3.86), and hierarchy/readability (3.91), respectively. These results support its ability to coordinate improvements across multiple production elements while maintaining temporally consistent outputs. Overall, the results demonstrate the method's strength in cinematic appearance enhancement and temporal consistency, the two complementary requirements of low-to-high-cost video editing.

4.3 Qualitative Comparison

We inspect representative results under matched inputs and instructions. Figure 5 presents two representative cases and exposes the trade-off between edit strength and source preservation. VACE often under-edits, while stronger baselines can alter faces or scene layout. FilmTogether introduces period-specific portrait cues and futuristic holographic panels and smoke while retaining identity, pose, and viewpoint across frames, yielding controlled enhancement rather than unconstrained redrawing. Additional examples of cinematic editing on everyday basketball and piano videos are provided in Appendix B.3.

Table 1: Quantitative comparison with 12 state-of-the-art video editing models. Best and second-best results are bold and underlined, respectively. All metrics: higher is better. Narr./Style/Lighting averages narrative/emotional fit, style/world consistency, and composition/lighting.

Method	CLIP-F	DINO-F	PickScore	VLM Aesthetic	Narr./Style/ Lighting	Color/ Texture	Hierarchy/ Readability
InsV2V (2024)	0.9857	0.9865	18.02	1.89	1.95	1.84	2.16
TokenFlow (2024)	0.9935	0.9956	18.85	2.98	2.99	2.96	3.18
VACE (2025)	0.9884	0.9896	18.09	2.31	2.34	2.26	2.53
InsViE (2025)	0.9918	0.9923	18.25	1.89	2.09	1.71	2.03
ICVE (2025)	0.9933	0.9976	17.94	2.53	2.55	2.52	2.84
Señorita (2025)	0.9953	<u>0.9980</u>	18.18	2.54	2.55	2.52	2.68
ReCo (2025)	<u>0.9962</u>	0.9977	<u>18.88</u>	3.44	3.47	3.40	3.43
Lucy-Edit (2025)	0.9946	0.9977	18.65	3.59	3.59	3.58	3.63
Bernini-R-14B (2026)	0.9903	0.9896	**19.05**	<u>3.77</u>	<u>3.80</u>	<u>3.74</u>	<u>3.80</u>
Ditto (2026)	0.9911	0.9913	18.83	3.22	3.26	3.18	3.42
Kiwi-Edit (2026)	0.9931	0.9958	17.79	2.28	2.28	2.28	2.31
SANA-Streaming (2026)	0.9937	0.9964	18.06	2.33	2.40	2.28	2.50
FilmTogether (Ours)	**0.9967**	**0.9985**	18.83	**3.87**	**3.88**	**3.86**	**3.91**

4.4 Dataset Ablation

The paired CineCraft-120K data improves VLM-Aesthetic from 3.22 to 3.78 for Ditto and from 3.66 to 3.87 for FilmTogether, while also improving temporal and fine-grained appearance scores (Table 2).

Table 2: Effectiveness of CineCraft-120K across two backbones. Best results within each backbone are bold. Green rows use CineCraft-120K. All metrics are higher-is-better.

Backbone	Dataset	CLIP-F	DINO-F	PickScore	VLM Aesthetic	Narr./Style/ Lighting	Color/ Texture	Hierarchy/ Readability
Ditto	Ditto Dataset	0.9911	0.9913	18.83	3.22	3.26	3.18	3.42
	CineCraft-120K	**0.9968**	**0.9985**	**18.87**	**3.78**	**3.80**	**3.76**	**3.82**
FilmTogether	Ditto Dataset	0.9917	0.9920	**19.04**	3.66	3.71	3.62	3.75
	CineCraft-120K	**0.9967**	**0.9985**	18.83	**3.87**	**3.88**	**3.86**	**3.91**

4.5 User Study

User Study Protocol. Ten participants compare *FilmTogether* with Ditto and Bernini-R-14B on ten randomly selected low-to-high editing cases. For each case, they view the source video, editing instruction, and three synchronized outputs with anonymized method labels. Case order is randomized across participants, and output positions vary across cases. Participants select one preferred output for each of six criteria: instruction following, cinematic enhancement, content preservation, motion fidelity, temporal consistency, and visual quality and realism, yielding 600 selections in total and 100 per criterion. We report selection rates for each criterion and their unweighted mean.

Human Preference Results. Figure 6 shows that *FilmTogether* achieves the highest observed selection rate across all six criteria, averaging 74.3%, compared with 21.2% for Bernini-R-14B and 4.5% for Ditto. It receives 72% of selections for cinematic enhancement, compared with 26% for Bernini-R-14B and 2% for Ditto. It also leads in content preservation (79%), motion fidelity (83%), temporal consistency (81%), and visual quality and realism (77%). The margin is smaller for instruction following (54% versus 43% for Bernini-R-14B). These results support FilmTogether's perceived ability to enhance cinematic appearance while retaining source content and motion on the evaluated cases.

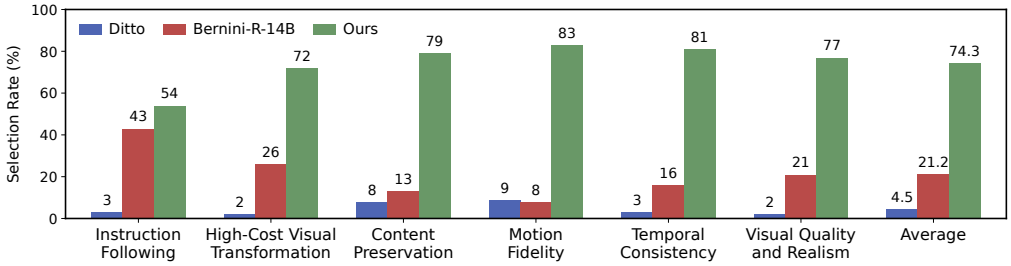


Figure 6: Human preference for low-to-high video editing. Values are three-way selection rates (%, ↑), with 100 responses per criterion from ten participants evaluating ten cases. Average is the unweighted mean over six criteria.

5 Conclusion

We present FilmTogether for transforming low-cost videos into cinematic outputs while preserving source content and motion. CineCraft-120K provides 23 diverse scenarios, aligned video pairs, and pair-aware editing instructions. Its dual-memory adapter coordinates semantic edits with layer-wise visual guidance. Quantitative results and human evaluation show improved cinematic appearance and temporal consistency, supporting accessible low-to-high video enhancement.

AI Use Statement

Generative AI tools assisted with manuscript language editing. The authors reviewed all assisted edits and take responsibility for the final content; scientific contributions, methods, results, and conclusions are the authors' own.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, et al. Scaling instruction-based video editing with a high-quality synthetic dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37971–37981, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Brandon Castellano. PySceneDetect. Software, n.d. URL `https://github.com/Breakthrough/PySceneDetect`.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23149–23160. IEEE, 2023.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22831–22840. IEEE, 2025.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *International Conference on Learning Representations*, volume 2024, pp. 16867–16879, 2024.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations*, volume 2024, pp. 52086–52106, 2024.

Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. Finevq: Fine-grained user generated content video quality assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3206–3217. IEEE, 2025.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xijie Huang, Chengming Xu, Donghao Luo, Xiaobin Hu, Peng Tang, Xu Peng, Jiangning Zhang, Chengjie Wang, and Yanwei Fu. Ffp-300k: Scaling first-frame propagation for generalizable video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23172–23181, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15908–15918. IEEE, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

Xinyao Liao, Xianfang Zeng, Ziye Song, Zhoujie Fu, Gang Yu, and Guosheng Lin. In-context learning with unpaired clips for instruction-based video editing. *arXiv preprint arXiv:2510.14648*, 2025.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance. *arXiv preprint arXiv:2603.02175*, 2026.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8599–8608. IEEE, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pp. 407–426. Springer, 2022.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15886–15896. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Bernini Team, Chenchen Liu, Junyi Chen, Lei Li, Lu Chi, Mingzhen Sun, Zhuoying Li, Yi Fu, Ruoyu Guo, Yiheng Wu, et al. Bernini: Latent semantic planning for video diffusion. *arXiv preprint arXiv:2605.22344*, 2026.

DecartAI Team. Lucy edit: Open-weight text-guided video editing, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7589–7599. IEEE, 2023.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16692–16701. IEEE, 2025b.

Zhongwei Zhang, Fuchen Long, Wei Li, Zhaofan Qiu, Wu Liu, Ting Yao, and Tao Mei. Region-constraint in-context generation for instructional video editing. *arXiv preprint arXiv:2512.17650*, 2025.

Yuyang Zhao, Yicheng Pan, Qiyuan He, Jincheng Yu, Junsong Chen, Tian Ye, Haozhe Liu, Enze Xie, and Song Han. Sana-streaming: Real-time streaming video editing with hybrid diffusion transformer. *arXiv preprint arXiv:2605.30409*, 2026.

Bojia Zi, Penghui Ruan, Marco Chen, Xianbiao Qi, Shaozhe Hao, Shihao Zhao, Youze Huang, Bin Liang, Rong Xiao, and Kam-Fai Wong. Señorita-2m: A high-quality instruction-based dataset for general video editing by video specialists. *Advances in Neural Information Processing Systems*, 38, 2026.

A Related Work

Video-to-video Datasets. Large-scale captioned video corpora and caption-transfer pipelines, including WebVid, VideoCC, Panda-70M, InternVid, and OpenVid-1M, support video-language pretraining but do not provide paired editing transformations (Bain et al., 2021; Nagrani et al., 2022; Chen et al., 2024; Wang et al., 2024; Nan et al., 2025). InsV2V synthesizes video pairs from related prompts (Cheng et al., 2024), whereas InsViE-1M, Ditto-1M, and FFP-300K use frame-guided editing and temporal propagation or synthesis (Wu et al., 2025b; Bai et al., 2026; Huang et al., 2026). Señorita-2M and OpenVE-3M employ specialist or category-specific pipelines for diverse editing operations (Zi et al., 2026; He et al., 2025). Appendix Table A1 summarizes the key characteristics of these approaches. Our CineCraft-120K specializes this strategy for cinematic enhancement, pairing curated high-production-value targets with synthesized low-cost counterparts and prompts grounded in observed differences. It focuses on coordinated upgrades to lighting, costumes, and materials while preserving scene structure and motion.

Instruction-based Video Editing. Diffusion-based video editors adapt image priors through one-shot tuning, zero-shot inversion and attention reuse, anchor-frame propagation, or flow-guided attention (Wu et al., 2023; Khachatryan et al., 2023; Qi et al., 2023; Ceylan et al., 2023; Liu et al., 2024; Cong et al., 2024). AnyV2V instead combines an image editor with image-to-video propagation in a tuning-free pipeline (Ku et al., 2024). Inversion-based methods such as TokenFlow (Geyer et al., 2024) reuse source features and inter-frame correspondences under target text guidance, while paired-data methods such as InsV2V (Cheng et al., 2024) and InsViE (Wu et al., 2025b) learn

Table A1: Comparison of video-editing datasets and the CineCraft-120K high-cost visual bank.

Dataset	Scale	Editing objective	Target provenance	Pair construction
InsV2V (Cheng et al., 2024)	68.4k	Object/scene/style edits	Synthetic	Video Prompt-to-Prompt
InsViE-1M (Wu et al., 2025b)	1.02M	Local/global edits	Generated/image-derived	Frame propagation + images
Señorita-2M (Zi et al., 2026)	2M	Edits + translation	Edited/original/derived	Specialists + transforms
Ditto-1M (Bai et al., 2026)	1M	Local/global edits	Generated	Frame edit + depth synthesis
OpenVE-3M (He et al., 2025)	3M	Multi-task editing	Generated/original	Category-specific
FFP-300K (Huang et al., 2026)	300k	Inpainting/stylization	Generated	Inpainting + stylization
CineCraft-120K (Ours)	124.9k	Cinematic enhancement	Curated high-cost videos	High-to-low synthesis

instruction-driven transformations. FFP-300K (Huang et al., 2026) uses an edited first frame to specify the target appearance. Ditto's Editto (Bai et al., 2026) bridges reference-guided training and instruction-driven inference by gradually removing this reference during training. OpenVE-Edit (He et al., 2025) further enriches conditioning through a video–instruction encoder and a task-aware MOE connector. Our dual-memory adapter augments text conditioning and refines layer-wise source hints within the existing editor, supporting cinematic instruction interpretation and low-cost appearance adaptation.

B Additional Experimental Details

B.1 Aesthetic Evaluator Details

Scoring rubric. The teacher and distilled evaluator use a four-level rubric for every aspect: 4 denotes an excellent or strongly successful result, 3 a good result with minor issues, 2 a weak result with clear issues, and 1 a failed result or one that harms the intended aesthetic. The five aspects are defined as follows.

1. **Narrative and emotional fit**: integration with scene mood, emotional tone, and visual context without distracting anomalies.

2. **Style and world consistency**: compatibility with the era, genre, environment, and overall visual language.

3. **Composition and lighting design**: composition, contrast, lighting hierarchy, lens design, and shot-level arrangement.

4. **Color and texture refinement**: color harmony, saturation control, material quality, and avoidance of cheap or generic filter looks.

5. **Visual hierarchy and readability**: clarity of the focal hierarchy and visibility of important content.

The holistic score is produced independently in $[1, 4]$ and is not defined as the arithmetic mean of the five aspect scores. Utility edits such as removal, cleanup, and denoising are not penalized for being visually subtle when the result is seamless and natural.

Student input and output. At evaluation time, frames are extracted from the edited video at 1 frame per second, resized to 640 pixels in width, and the first two available frames are used. The source video and high-cost target are excluded. The instruction is supplied only as weak context for the intended visual direction. The evaluator must emit a valid JSON object containing the five integer aspect scores, a floating-point holistic score, an uncertainty flag, and a one-sentence rationale. At test time we disable sampling and average each numeric field over the 50 examples.

B.2 METHOD-LEVEL MEMORY ABLATION

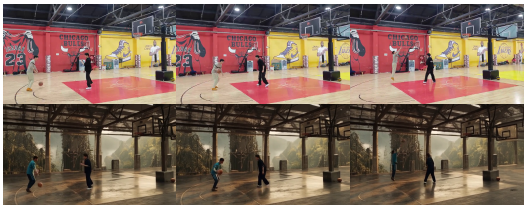
The component-removal comparison measures CLIP-F and DINO-F for the full model and three memory ablations. The scores are tightly clustered; the values are reported as measured, without combining checkpoints (Table A2).

Table A2: Memory ablation on CineCraft-Bench. Green highlights FilmTogether.

Configuration	CLIP-F	DINO-F
w/o Text Memory	0.9935	0.9939
w/o Hint Memory	0.9933	0.9938
w/o both memories	0.9933	0.9939
FilmTogether	**0.9934**	**0.9939**

B.3 CINEMATIC EDITING OF EVERYDAY VIDEOS

We further illustrate FilmTogether on everyday basketball and piano videos. Figure A1 shows how the model transforms ordinary environments into cinematic settings through coordinated changes to architecture, lighting, clothing, and materials. The basketball example introduces a mountainous setting with wooden architecture and atmospheric haze, while the piano example transforms a casual venue into an elegant lounge with chandeliers and warm lighting. Three corresponding frames illustrate these appearance changes over time.



Place the game in a cinematic mountain setting with traditional wooden architecture and atmospheric haze. Dress the players in Chinese-style outfits. Preserve identities and actions.



Create a luxury piano lounge with chandeliers and warm lighting; add black attire and a polished dark-wood piano. Preserve faces, playing motions, keyboard, and camera movement.

Figure A1: Cinematic editing of everyday videos. Basketball and piano examples show three temporally corresponding frames from the input videos (top) and FilmTogether outputs (bottom). Editing prompts are provided below each example, with key transformation attributes highlighted in red.

B.4 MEMORY LEARNING DYNAMICS

The trajectories diagnose code differentiation and parameter adaptation rather than generation quality. They complement the performance-based ablations in the main paper by showing how the two memory banks evolve during continuation training.

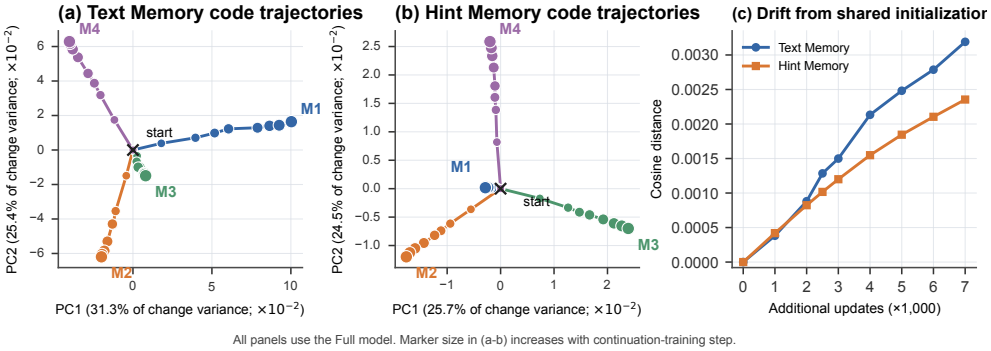


Figure A2: Learning dynamics of the dual-memory adapter. (a–b) PCA trajectories of the four Text and Hint Memory codes during continuation training of the Full model; marker size increases with optimizer updates. (c) Cosine distance of each complete codebook from the shared Table-1 initialization.