Eurographics Symposium on Rendering (2026)
I. Gkioulekas and A. Jarabo (Editors)

CoherentEdit: Unified Multimodal Understanding and Implicit Lighting Conditioned Diffusion for Controllable Video Editing

Guojun Lei[1], Zheng Dong[2], Hongbing Yang, Hong Li[3], Yikai Wang, Chi Wang[1], and Weiwei Xu[1]

[1]Zhejiang University, China
[2]Zhejiang Sci-Tech University, China
[3]Beihang University, China

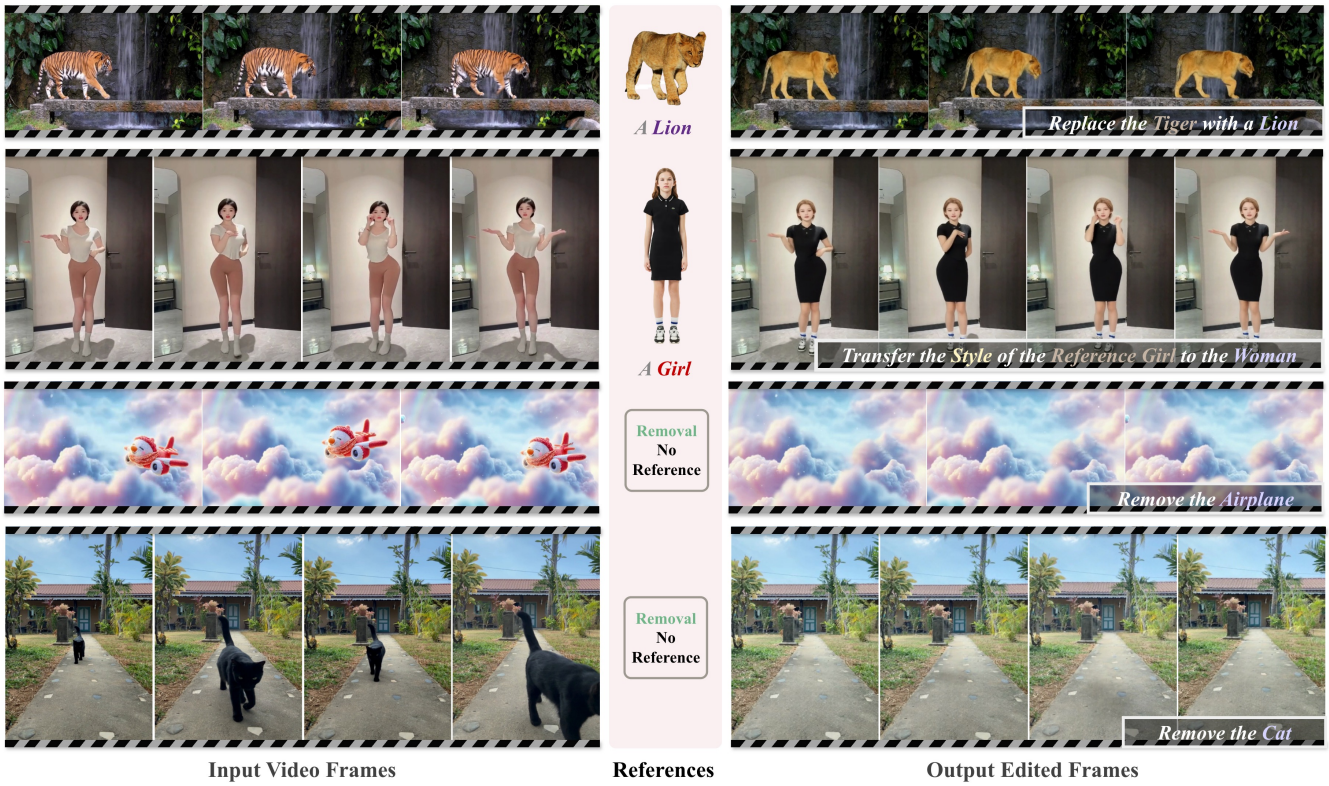

Figure 1: *We propose CoherentEdit, which unifies video understanding and generation, leveraging text prompts and optional reference images to enable diverse video editing tasks, such as object removal, replacement, and style transfer.*

Abstract

Recent progress in video generation has intensified demand for controllable video editing. Vision-language models have demonstrated strong capabilities in video understanding, yet they appear to offer limited support for video generation and editing. We present CoherentEdit, a two-stage framework for unified multimodal understanding and video editing. We explore how to leverage the powerful video understanding capabilities of Vision-Language Models (VLMs) to better serve video generation and editing. In the first stage, the VLM model interprets the user instruction and the source video to derive a unified multimodal embedding and we explored how to push the VLM output from text space to multimodal space. In the second stage, a diffusion-based generation network performs controllable video synthesis. To better preserve lighting coherence, CoherentEdit extracts an implicit lighting representation from the source video via a VQ-VAE encoder, and injects this representation into the generation process to promote consistent illumination and shadowing across edited regions and unchanged context. Experiments demonstrate that our method outperforms the state-of-the-art approaches. For more details and videos, please refer to the anonymous webpage: https://anonymous-name-paper.github.io/anonymous-project/.

DOI: 10.2312/sr.20261020

1. Introduction

Unifying multimodal understanding and generation has become a central goal in the development of multimodal large language models (MLLMs) [ZYD*24, YFZ*24]. Recent advances, such as Gemini 2.0 Flash [TAB*23], GPT-5 [Ope25], Bagel [DZL*25], and OmniGen2 [WZY*25], demonstrate that a single framework can jointly handle perception, reasoning, generation, and editing across modalities. These achievements mark a shift from task-specific networks to multimodal models that learn unified representations for visual-textual understanding and synthesis.

However, achieving such unification in the video domain remains challenging. Most recent work on video understanding uses autoregressive models (e.g., Flamingo [ADL*22], Qwen2.5-VL [BCL*25]), whereas video generation relies on diffusion-based models (e.g., Sora [LZL*24b], Hunyuan-Video [KTZ*24]). A major gap exists because the former use discrete tokens for temporal reasoning, while the latter operate on continuous latent representations. Aligning the feature spaces and training strategies of these two model families is non-trivial. Moreover, effective unification requires large-scale paired video-editing data, which is far more difficult to obtain than image editing data, further limiting end-to-end training. Although HunyuanCustom [HYZ*25], Uni-Video [WLY*25], and OmniVideo [TYQ*25] incorporate Vision-Language Models (VLMs), they commonly use VLM hidden states as a replacement for the text encoder, often relying on the final-token representation. Our analysis shows that this final-token hidden state is frequently dominated by sentence-ending or formatting information and is therefore a weak condition for video editing. In summary, current VLMs primarily support multimodal inputs but are not explicitly trained to produce multimodal editing conditions. Their outputs remain biased toward the language space, which makes direct coupling with continuous-latent video diffusion models suboptimal.

To address these issues, we propose CoherentEdit, a two-stage framework that unifies video understanding and generation within a coherent pipeline for diverse editing tasks. In the first stage, CoherentEdit employs a video understanding network to encode semantic instructions and global video context into latent features. In the second stage, a diffusion-based generator performs instruction- and reference-guided video synthesis conditioned on these aligned latent features. To bridge the inherent gap between discrete tokens in autoregressive models and continuous latents in diffusion models, we design task-specific alignment tokens and train them with our self-supervised editing tasks. These tokens provide hidden states that are optimized for video-text interaction rather than sentence completion, thereby pushing the VLM outputs from the original text space toward a multimodal conditioning space, as illustrated in Fig. 2. Furthermore, to mitigate the scarcity of labeled video-editing data, we adopt a self-supervised learning strategy that automatically generates high-quality pseudo-editing pairs to facilitate the training of our two-stage framework.

We also observe that many edited videos contain a visible mismatch between edited foreground objects and their backgrounds, particularly in terms of lighting and shadows. To further enhance visual coherence, we extract quantized implicit illumination features from the source video via VQ-VAE [VDOV*17] and feed them into the video generation stage, resulting in more consistent lighting and shadows across frames.

Figure 2: *Token-space alignment visualization. Hidden states are collected from validation samples and projected to two dimensions using PCA. Blue points denote learned task-alignment tokens, while the remaining points denote ordinary language-token hidden states. The visualization shows that task-alignment tokens form editing-oriented conditions that are separated from sentence-ending language tokens.*

We validate the effectiveness of our unified framework on various video editing tasks, including *object removal, insertion, replacement, and attribute modification*. Our results demonstrate strong qualitative and quantitative performance across multiple video editing benchmark metrics.

Our contributions are summarized as follows:

- We propose *CoherentEdit*, a unified framework for video understanding and video editing. Rather than simply replacing the text encoder with a Vision-Language Model (VLM), CoherentEdit aligns VLM output features with diffusion-model conditioning features to better exploit VLM video reasoning for controllable editing.
- We mitigate the gap between discrete VLM tokens and continuous diffusion embeddings through self-supervised training with task-specific alignment tokens, avoiding reliance on large-scale manually annotated video-editing pairs.
- We model environment illumination via VQ-VAE and implicitly transfer it to the edited output, ensuring photorealistic lighting and consistent shading. We validate our approach and design through comprehensive benchmarks across multiple video editing tasks.

2. Related Works

2.1. VLM-based Video Understanding

Video understanding seeks to analyze video content and reason about the temporal relationships [ZAOT18]. Early VLMs established that LLM can ground visual inputs and perform open-ended reasoning [WYM*22, LLSH23, LLWL23]. Among these works, Flamingo [ADL*22] and Video-LLaMA [ZLB23] have demonstrated effectiveness in processing video content for interpretive tasks, including visual question answering [KSX*25, LCLY24, HCS*25] and summarization [HTXL25, TBX*25], by capturing both global semantics and temporal information. Recent video LLMs push toward longer sequences and finer tem-

poral resolution [QDZ*24, WHH*24, LWJ24, SCW*24]. Video-ChatGPT [MRKK24] targets detailed conversational understanding of videos, while Qwen2.5-VL [BCL*25] provides video perception by capturing long-range temporal dependencies and fine-grained dynamics across entire sequences. MovieChat [SCW*24] introduces sparse memory to manage long-range sequences and releases a long-video benchmark. However, while the methods mentioned perform well in video understanding, they cannot drive generative control or leverage MLLM reasoning to guide the video editing process.

2.2. Controllable Video Editing

Video editing has progressed from early optimization pipelines based on optical flow [XLZL19, LLQ*22] and per-frame masks [ZLCL23] to deep generative methods that enable more realistic and controllable edits. GAN-based methods [XAH22, FSX*23] improved visual fidelity but often suffered from unstable training and weak temporal consistency. With the rise of diffusion models [CHIS23], video editing [LZL*24a, QCZ*23] achieves markedable improvements in quality and controllability. Methods such as Tune-A-Video [WGW*23] and Control-A-Video [CWX*23] adapt image diffusion backbones with temporal modules to achieve content-consistent edits. Zero-shot approaches such as Text2Video-Zero [KMT*23] further reduce the per-video tuning by reprogramming cross-frame attention. Video-P2P [LZL*24a] extends this line by employing cross-attention control and a decoupled-guidance strategy after video inversion. More recently, Diffusion Transformers (DiTs) [LDH*21, CLT*22, YZLL23, PX23] have become strong backbones for long-range temporal modeling and high-fidelity video generation (e.g., CogVideoX [YTZ*24], Hunyuan-Video [KTZ*24] and Sora [LZL*24b]). Within controllable editing, CCEdit [FWW*24] separates structure and appearance control based on diffusion and ControlNet [ZRA23], and adds temporal layers to improve motion consistency. UniAnimate [WZG*25] maps appearance and driving pose to a shared space and adopts state-space temporal modeling to produce identity-consistent videos. Despite these gains, many pipelines remain task-specific (one model per edit type) [BTOAF*22, CWX*23, ZLCL23] design and lack physically consistent modeling (e.g., lighting and shadows) [WGW*23, KMT*23], thus they cannot achieve unified and realistic video editing. In contrast, we unify video understanding and video generation editing via aligned multimodal space and integrating implicit illumination, which helps achieve scene-aware, consistent, and controllable video editing.

3. Method

Our proposed CoherentEdit aims to integrate video understanding and generation into a unified two-stage pipeline, as illustrated in Fig. 3. Firstly, a VLM with strong video reasoning capabilities encodes both the input video and textual prompts into unified multimodal embeddings. Secondly, these embeddings, along with detailed reference features, are used to guide the generation process for the diffusion model.

3.1. Unified Video Understanding

We employ QwenVL-2.5-7B [BCL*25] as our vision-language model (VLM). However, we observe that while VLMs possess strong visual understanding capabilities, their outputs consist of discrete text tokens, which cannot be directly utilized for the continuous and spatially coherent requirements of video editing tasks.

To bridge this gap, we introduce a dedicated adaptation and training strategy. Specifically, standard VLMs are typically trained on video captioning tasks—generating descriptive text from video content. However, when presented with editing instructions such as "erase the airplane in this video", the model lacks clear guidance on what to output. This is due to two key challenges: (1) the absence of ground-truth edited videos as supervision signals, and (2) the difficulty of representing complex video edits using only discrete textual tokens.

To address these issues, we design a set of special task-specific tokens: $< /generate >$, $< /eraser >$, $< /add >$. When the input prompt involves a video editing operation, we use the corresponding special token as the target output during training, to push the model's outputs from the original text space into multimodal space shown in Fig. 5.

Accordingly, we restructure the VLM's training data into four distinct components aligned with these tasks:

- $< /generate >$: For standard video generation or reconstruction tasks. The prompt is formulated as "According to this video, generate a video similar to this video".
- $< /erase >$: We randomly erase certain regions (e.g., objects) in the video frames and pair them with the instruction "According to this video, identify what is missing or problematic and try to fix it". This encourages the model to recognize incompleteness and infer appropriate restoration or editing behavior.
- $< /add >$: We synthetically insert animals or human figures into random regions of the video and use the same instruction "According to this video, identify what is missing or problematic and try to fix it". This trains the model to detect unnaturalness or anomalies and reason about plausible additions.
- Captioning (original): We retain the original captioning data without modification, preserving the VLM's general video comprehension ability.

3.2. Unified Video Generation for Editing

Video diffusion model can be defined as follows: Assuming $x_0 \in R^{f\times h\times w\times c}$ represent a video latent with f frames, each with dimensions $h \times w$. The forward diffusion process is defined as a Markov chain which adds Gaussian noise to the original video step by step:

$$x_t = \sqrt{\bar{a}_t}x_{t-1} + \sqrt{(1-\bar{a}_t)}\varepsilon, \varepsilon \sim N(0,1), \quad (1)$$

where $t \in 1,...,T$ denotes the diffusion step, $\bar{a}_t$ is a noise scheduler and ε is sampled from a standard Gaussian noise. The model learns to reverse this process, estimating $p(x_{t-1}|x_t)$. In this paper, we adopt a diffusion transformer (DiT) architecture for noise estimation. Overall, DiT-based models have demonstrated superior stability and generation quality compared to Unet-Based architectures.

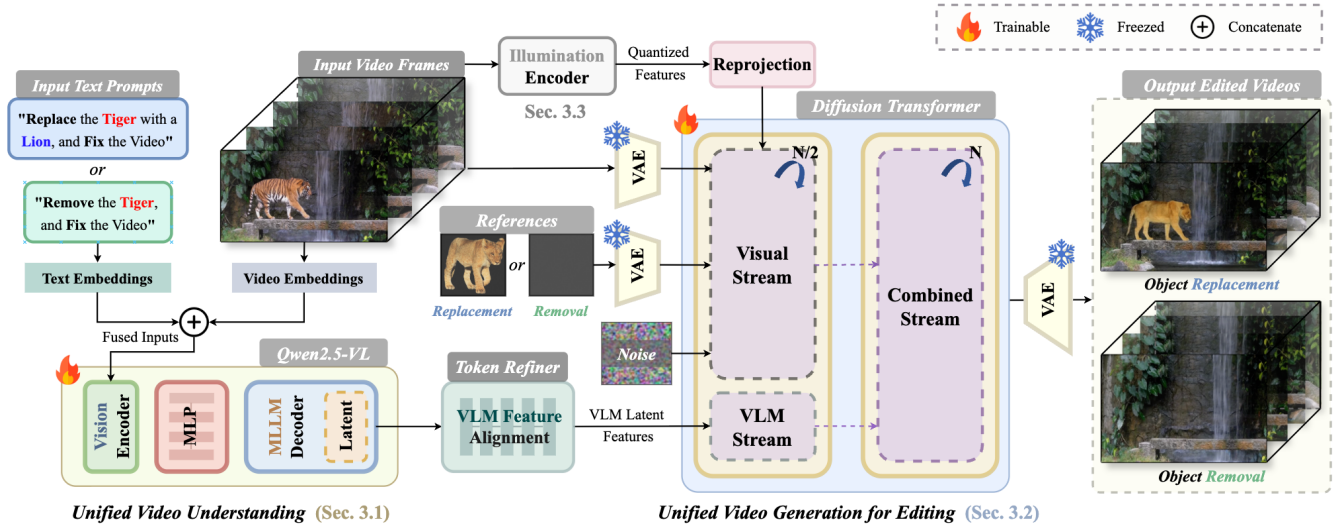

Figure 3: *The pipeline of our CoherentEdit: 1) a VLM first encodes video frames and text prompts into unified multimodal embedding; 2) A DiT-based diffusion model receives detailed reference features and multimodal latent features to produce final results with coherent appearance and realistic lighting under the projection of implicit environment lighting.*

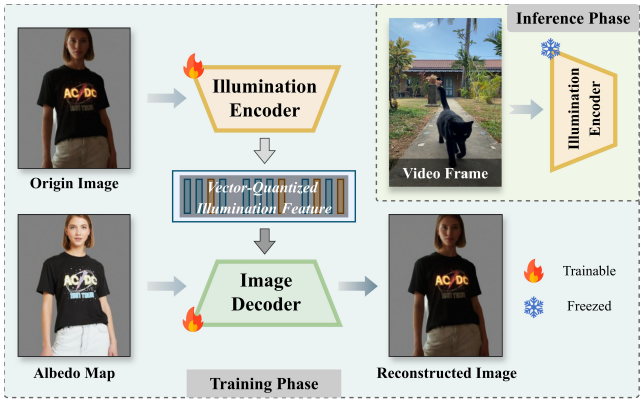

Figure 4: *Architecture of our VQ-VAE used for implicit illumination latent encoding and reconstructing images under the albedo maps.*

Training large models like Huanyuan-video [KTZ*24] for video editing is often hampered by a critical data bottleneck. An end-to-end approach typically requires large-scale paired data, which are exceptionally rare and difficult to collect. Manually curating or synthetically generating paired video editing data is a significant challenge.

The foundation of our approach is a custom dataset we constructed from the WebVid [BNVZ21a] and OpenVid [NXZ*24] using a self-supervised way. We synthetically generate training examples by defining three distinct editing tasks aligned with the training of the VLM model:

- **Object Erase** $< /erase >$**:** We randomly mask certain regions (e.g., objects) in the original video frames and train the diffusion model to restore the original video.
- **Object Insert** $< /add >$**:** We insert objects (animals or humans) into random locations within the video and train the model to restore them.
- **Reconstruction** $< /generate >$**:** The original, unmodified video serves as both the input and the target for high-fidelity video generation.

In all three scenarios, the unmodified input video (before erasement or insertion) serves as the ground-truth during training. This self-supervised setup allows us to generate realistic editing supervision without requiring manually annotated or edited video pairs shown in Fig. 5.

Our video generation backbone is built upon Hunyuan-Video [KTZ*24]. Our approach differs fundamentally from Omni-Video [omn26] and UniVideo [WLY*25], which directly feed the hidden state of the last token from the final transformer block into the subsequent generator. Through careful analysis, we observe that while Vision-Language Models (VLMs) possess multimodal input capabilities, they inherently lack multimodal output capabilities. To bridge this gap, we deliberately design task-aligned multimodal control tokens (such as $< /add >$,$< /erase >$, et al.)to steer the VLM's output from the textual domain into a multimodal representation space, as illustrated in Fig.2. Consequently, we discard the original text stream and replace it entirely with VLM stream. The VLM model is designed to capture the correlation between user prompts and corresponding videos, which outputs comprehensive video embeddings serving as unified multimodal embeddings for the video generation. As illustrated in our pipeline (Fig. 3), in the visual stream, we incorporate the original background video and the reference image to provide detailed spatial and image-level guidance, ensuring that the generated content aligns well with both the detailed global scene structure and the desired new reference image. The architecture is built upon two complementary transformer block designs. The first part, maintains separate processing

pathways for visual embedding and VLM embedding throughout the first 20 layers of the network. Crucially, these separate streams converge through a joint attention mechanism where the query, key, and value matrices from both modalities are concatenated before computing the attention weights. After each visual stream block, we integrate the implicit environment light through separate attention to merge the environment light features, referring to section3.3, which uses the illumination latents z_l to query video latents z_v and add a residual connection before the next visual stream block, as:

$$z_v' = \text{Attention}(\mathrm{Q}_l,\ \mathrm{K}_v,\ \mathrm{V}_v) = \text{Softmax}\left(\mathrm{QK}^{\mathrm{T}}\right)\mathrm{V}, \tag{2}$$

where $Q_l = W^Q z_l$,$K_v = W^K z_v$,$V_v = W^V z_v$. After the double processing stage, the visual embedding and VLM embedding are concatenated into a unified representation that passes through the combined stream to do full attention comprising 40-layer blocks. To prevent gradient explosion caused by an excessive number of training parameters, we adopted a separate training approach shown in section4.1.

3.3. Implicit Video Illumination Transfer

To alleviate visual inconsistency between edited regions and their background, an ideal solution would estimate scene lighting, object geometry, and material properties jointly. Recent inverse-rendering methods [BBJ*21, ZSH*22, LGL*25] optimize lighting and scene parameters with photometric losses, but they often require dense multi-view observations or known geometry. Relighting methods such as IC-Light [ZRA25] and RelightVid [FSZ*25] focus on relighting images or videos, but they are not designed to extract a reusable illumination condition from an arbitrary source video for editing. Accordingly, we avoid explicit physical light reconstruction and instead learn an implicit illumination representation from controlled-lighting supervision. We assume that the source-video illumination can be represented as a smooth spatio-temporal factor that should be preserved around the edited foreground and its nearby background. This representation is injected into the generator as an auxiliary condition rather than used as a complete physical lighting model.

We adopt a Transformer-based VQ-VAE [OVK] for implicit illumination encoding, as shown in Fig. 4. Given an illuminated image I and its albedo image A, the illumination encoder E_l produces continuous light tokens $h = E_l(I)$. A vector-quantization codebook $\mathcal{C} = \{e_k\}_{k=1}^K$ maps each token to its nearest codeword:

$$z_l = \text{VQ}(h) = e_{k^*}, \quad k^* = \arg\min_k \|h - e_k\|_2. \tag{3}$$

In our implementation, the codebook contains $K = 1024$ entries with dimension $d = 256$, and each image is represented by $n = 16$ quantized illumination tokens. The decoder D reconstructs the illuminated image from the albedo image and the quantized light tokens, $\hat{I} = D(A, z_l)$. The VQ-VAE is trained with reconstruction, perceptual, codebook, and commitment losses:

$$\mathcal{L}_{\text{vq}} = \|I - \hat{I}\|_1 + \lambda_p \mathcal{L}_{\text{perc}}(I, \hat{I}) + \|\,\text{sg}(h) - z_l\|_2^2 + \beta\|h - \text{sg}(z_l)\|_2^2, \tag{4}$$

where $\text{sg}(\cdot)$ denotes stop-gradient, λ_p controls the perceptual term, and β is the commitment weight. The compact codebook does not formally guarantee perfect illumination disentanglement; rather, under controlled-lighting supervision it empirically encourages the latent tokens to capture lighting-related factors instead of high-level semantic content.

At inference, we extract $\mathbf{z}_l(t)$ from the raw video frames and inject it into the generation network through the attention module described above. We further analyze the light-latent attention response during generation and observe stronger responses around the edited foreground subject and nearby regions, while unchanged background areas receive weaker responses. This behavior suggests that the illumination latent is mainly activated where foreground-background lighting harmonization is needed, helping maintain coherent shading, color temperature, and shadow consistency around edited objects.

4. Experiment

To better demonstrate the strengths of our approach, we conduct experiments to validate object editing and the generated video quality across various datasets and video generation benchmarks.

4.1. Implementation Details

We train the whole pipeline in three stages: 1) the VLM, 2) the implicit-illumination VQ-VAE, and 3) the video editing backbone. First, for the VLM component, we adopt Qwen2.5-VL [BCL*25] as the foundation model and train the task-alignment tokens on WebVid [BNVZ21b] with a learning rate of 5×10^{-5} for 10,000 iterations. Second, to train the implicit-illumination VQ-VAE, we construct a dataset of 30,000 controlled-lighting samples, including Internet images and samples generated with DiffusionLight [PCS*23], RelightVid [FSZ*25], and IC-Light [ZRA25]. The VQ-VAE is trained with a learning rate of 1×10^{-4} for 10,000 iterations. Finally, for the video editing backbone, we use the Hunyuan-Video architecture, which jointly processes VLM alignment-token embeddings, source-video features, reference-image features, and implicit illumination latents. Video generation is trained on WebVid and OpenVid [NXZ*24] with a learning rate of 10^{-4} for 20,000 steps. Relative to the Hunyuan-Video backbone, our added editing modules increase the parameter count by about 200M. We use AdamW for all stages; full training takes about four days on 16× NVIDIA A800-80GB GPUs, and inference for a single edited video takes about five minutes on 4× NVIDIA A800 GPUs. The code, trained weights, and data construction scripts will be released after publication for reproducibility.

4.2. Comparison with Previous Methods

Our training data is exclusively sourced from OpenVid and WebVid datasets. To ensure experimental fairness, we have not used any closed-source or internally produced data. Moreover, all evaluation data is drawn from open-source datasets that were not included in the training phase. Following Propainting [ZLCL23], we conduct experiments on both YouTube-VOS [XYF*18] and DAVIS [PPTM*16] to evaluate the performance of our method for object erasement from videos. We compared with Propainting, DiffuEraser [LXRB25] and HunyuanCustom [HYZ*25] using multiple metrics, including LPIPS, PSNR, FVD to assess the temporal consistency and smoothness of the resulting video sequences.

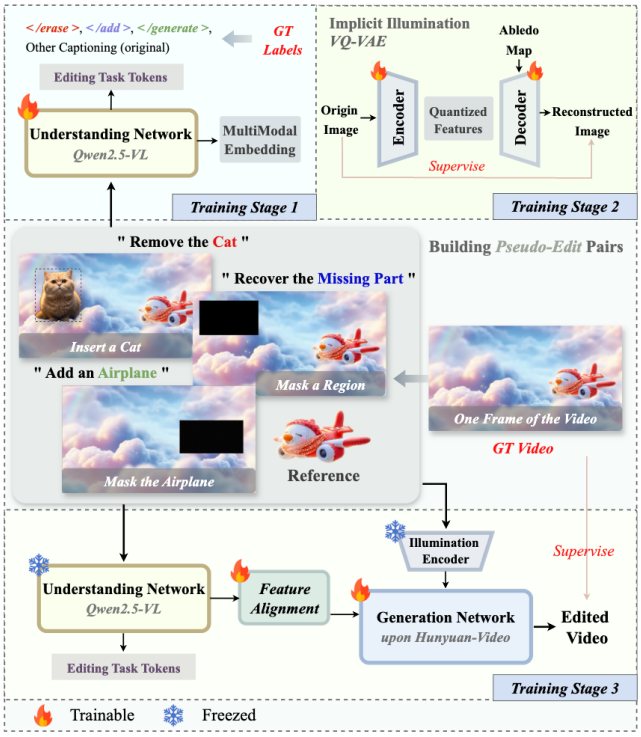

Figure 5: *Data synthesis pipeline and self-supervised training: 1) we randomly insert objects into videos or mask regions to create corrupted inputs; 2) the video understanding model learns task-alignment tokens, while the video generation model reconstructs the clean source video using VLM embeddings and implicit illumination features.*

The detailed results are presented in the Tab. 2 and Fig. 8. Benefit from the unified design of our framework and the implicit illumination transfer, the edited video after object erasement appears significantly more realistic, with fewer artifacts of inconsistent lighting and shadows, as shown in Fig. 8 and Fig. 6. Our method also achieves superior performance across multiple metrics, as shown in Tab. 2.

To further evaluate the generalization ability of our approach, we additionally test the quality of human-centric editing and motion control on the TikTok [JP21] and UBC fashion video datasets [ZSZS19].

To validate the overall video quality and temporal consistency, we also adopt some metrics from VBench [HHY*24] like Subject Consistency, Background Consistency, Motion Smoothness, Aesthetic Quality, and Dynamic Degree, and we compare our method with state-of-the-art character animation and video editing methods, including HunyuanCustom (HunyuanCus) [HYZ*25], VACE [JHM*25], Omni-Video [TYQ*25] (Wan2.1 [WWBA25]-based video generation model), , as shown in Tab. 1. We present the qualitative results in Fig. 7. Omni-Video only supports text-guided video editing at a coarse semantic level. HunyuanCustom produces relatively good results but tends to exhibit distortions, particularly in details such as hands, and also requires high alignment between the reference image and reference video. VACE could replace objects well but struggles to control corresponding poses. Other approaches also suffer from similar issues. In contrast, our method achieves superior performance in terms of lighting consistency and subject fidelity, even when the reference image and reference video are not well aligned, as shown in Fig. 7. This demonstrates the robustness and generalization ability of our approach in video generation and various editing tasks. See our appendix for more information.

4.3. Ablation Study and Analysis

To validate the effectiveness of our overall design, we conducted a series of ablation studies on the TikTok dataset. The experimental details are as follows:

VLM Performance. We first investigate the role of the VLM by replacing it with a standard T5 text encoder [RSR*23], which is widely adopted in video generation models. This variant corresponds to the first row of Table 3 and Fig. 9. For simple textual descriptions, VLMs and T5 provide comparable coarse semantics. However, in complex scenarios such as object editing, background completion, and artifact removal, the model needs to reason over the source video and the desired edited state. The VLM provides more useful video-conditioned semantics for such cases, as illustrated in Fig. 9.

VLM Alignment. We further validate the effectiveness of aligning VLM outputs with diffusion-model inputs. Taking object erasure as an example, directly decoding a frozen VLM produces a fluent textual instruction, as shown in Fig. 9, but such text-only output is not the condition needed by a diffusion generator. The generator requires grounded multimodal features that describe how the edited content should appear in the scene. As shown in Table 3, using a pre-trained VLM without task-specific alignment still leads to a performance decline. In contrast, our aligned task-token training produces more suitable editing embeddings and yields more accurate and detailed generated videos.

Implicit Illumination Transfer. We evaluate the impact of removing implicit illumination transfer. Quantitative ablations are reported in Table 3. As shown in Fig. 8 and Fig. 9, existing approaches and our ablated variants tend to leave visible artifacts or inconsistent shading around the edited region. In contrast, the full model better harmonizes foreground and background lighting, leading to improved visual quality and temporal consistency.

5. Conclusion

In this paper, we present a unified video editing pipeline, named CoherentEdit, to effectively integrate video understanding and generation. Our approach unfolds in two stages: first, a VLM encodes the input video and editing instruction into a comprehensive embedding; second, a diffusion-based generator leverages this embedding to produce controllable, semantically consistent edits, and we prove that task-specific tokens can better bridge the gap between discrete tokens from VLM models and continuous embeddings from diffusion models. We further incorporate implicit illumination features extracted from the source video to maintain lighting

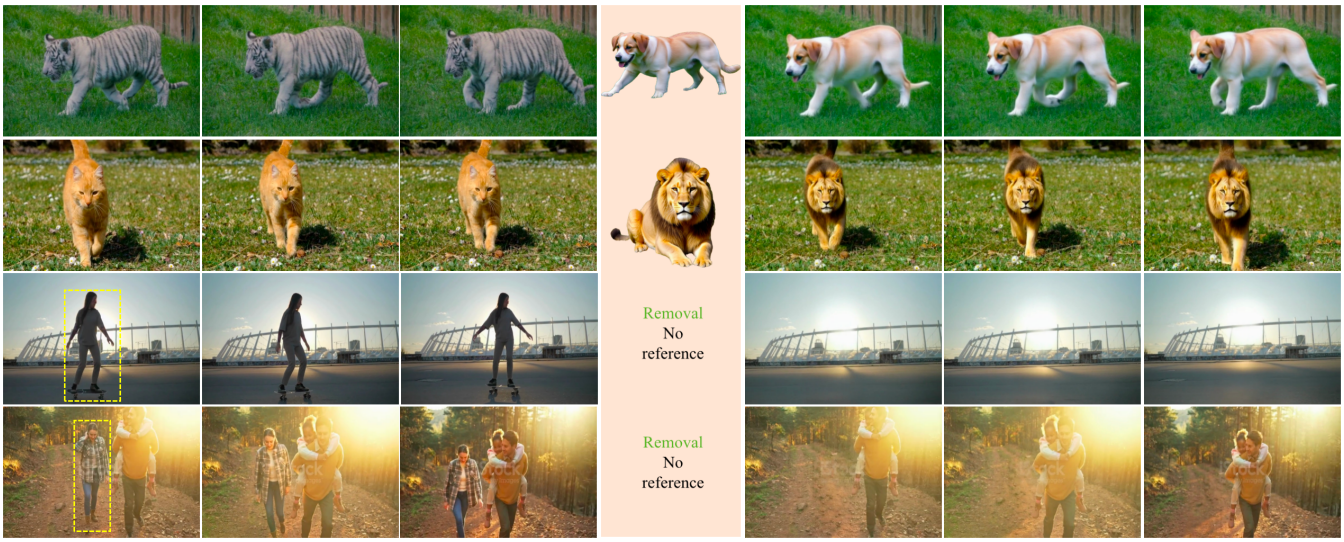

Figure 6: *Our method enables efficient video editing, such as object replacement or removal in diverse scene settings with single or multiple subjects, preserving visual consistency and temporal smoothness after editing.*

Table 1: *Quantitative comparisons of video consistency. SubC: Subject Consistency; BkgC: Background Consistency; MoS: Motion Smoothness; AesQ: Aesthetic Quality; DyaD: Dynamic Degree.*

Method	TikTok [JP21]					UBC [ZSZS19]				
	SubC↑	BkgC↑	MoS↑	AesQ↑	DyaD↑	SubC↑	BkgC↑	MoS↑	AesQ↑	DyaD↑
RF-Editor(hunyuan) [WPQ*24]	0.882	0.911	0.893	0.791	0.714	0.796	0.813	0.885	0.762	0.802
Omni-Video(wan2.1) [TYQ*25]	0.834	0.769	0.711	0.462	0.625	0.825	0.791	0.901	0.525	0.703
CogVideoX+Image [YTZ*24]	0.927	0.904	0.926	0.557	0.842	0.911	0.917	0.911	0.663	0.902
HunyuanVideo [KTZ*24]	0.914	0.851	0.789	0.624	0.752	0.909	0.872	0.741	0.625	0.812
Mofa-Video [NCW*24]	0.923	0.917	0.962	0.593	0.837	0.923	0.879	0.957	0.612	0.829
VACE(wan2.1) [JHM*25]	0.935	0.922	0.933	0.776	0.891	0.931	0.902	0.927	0.799	0.845
HunyuanCustom [HYZ*25]	0.937	0.898	0.902	0.801	0.891	0.924	0.907	0.941	0.795	0.902
Ours	0.941	0.922	0.967	0.812	0.899	0.937	0.917	0.962	0.803	0.913

Table 2: *Quantitative comparisons of video quality.*

Method	YouTube-VOS [XYF*18]			DAVIS [PPTM*16]		
	LPIPS↓	PSNR↑	FVD↓	LPIPS↓	PSNR↑	FVD↓
ProPainter [ZLCL23]	0.198	34.43	367	0.187	34.47	348
DiffuEraser [LXRB25]	0.178	32.13	326	0.171	32.16	343
HunyuanCustom [HYZ*25]	0.163	29.87	397	0.175	28.36	382
Ours	0.156	34.51	319	0.146	35.11	337

Table 3: *Ablation study with designs of our CoherentEdit.*

Method	Visual Quality			Video Smoothness	
	LPIPS ↓	PSNR ↑	FVD ↓	SubC ↑	BkgC ↑
w/o VLM Embedding	0.356	23.16	469	0.734	0.854
w/o VLM Alignment	0.215	24.11	391	0.893	0.832
w/o implicit Light	0.279	24.13	376	0.856	0.812
Full Model	0.143	27.13	336	0.941	0.922

and shading realism across edited content. Our framework shows strong potential for a wide range of editing tasks and practical applications.

6. Impact Statement

CoherentEdit has the potential to meaningfully advance the field of controllable video editing by bridging the long-standing gap between multimodal video understanding and generative editing models. By demonstrating how a Vision-Language Model's rich semantic understanding can be pushed beyond text-only outputs into a unified multimodal embedding space, this work opens a pathway toward more interpretable, instruction-driven video editing systems that operate with greater fidelity to user intent and could facilitate future research on multimodal alignment, controllable synthesis, and human-centered creative tools.

Figure 7: *Qualitative comparisons of subject replacement. Our method can generate motions and high-fidelity details that are consistent with the input video and the reference image, respectively.*

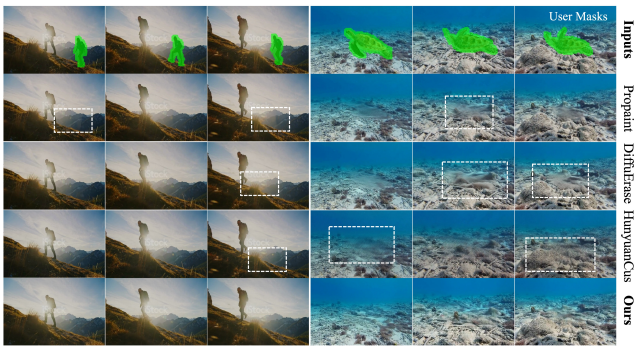

Figure 8: *Qualitative comparisons of object erasement.*

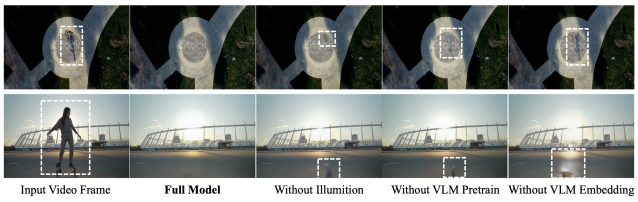

Figure 9: *Ablation study for object erasement task. Freezed VLM's output of the second row: To erase the human from the video sequence while preserving background continuity and consistent motion throughout the video, follow these steps for each frame: Remove the person and their shadow from the top-right section of the circular path. Verify that all frames now show the empty circular path with consistent lighting and no traces of the person or shadow.*

References

[ADL*22] ALAYRAC J.-B., DONAHUE J., LUC P., MIECH A., BARR I., HASSON Y., LENC K., MENSCH A., MILLICAN K., REYNOLDS M., ET AL.: Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems 35* (2022), 23716–23736. 2

[BBJ*21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCH H. P.: Nerd: Neural reflectance decomposition from image collections. In *IEEE International Conference on Computer Vision (ICCV)* (2021). 5

[BCL*25] BAI S., CHEN K., LIU X., WANG J., GE W., SONG S., DANG K., WANG P., WANG S., TANG J., ZHONG H., ZHU Y., YANG M., LI Z., WAN J., WANG P., DING W., FU Z., XU Y., YE J., ZHANG X., XIE T., CHENG Z., ZHANG H., YANG Z., XU H., LIN J.: Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025). 2, 3, 5

[BNVZ21a] BAIN M., NAGRANI A., VAROL G., ZISSERMAN A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision* (2021). 4

[BNVZ21b] BAIN M., NAGRANI A., VAROL G., ZISSERMAN A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 1728–1738. 5

[BTOAF*22] 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* (2022), Springer, pp. 707–723. 3

[CHIS23] CROITORU F.-A., HONDRU V., IONESCU R. T., SHAH M.: Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence 45*, 9 (2023), 10850–10869. 3

[CLT*22] CAI J., LI C., TAO X., YUAN C., TAI Y.-W.: Devit: Deformed vision transformers in video inpainting. In *Proceedings of the 30th ACM international conference on multimedia* (2022), pp. 779–789. 3

[CWX*23] CHEN W., WU J., XIE P., WU H., LI J., XIA X., XIAO X., LIN L.: Control-a-video: Controllable text-to-video generation with diffusion models. *CoRR* (2023). 3

[DZL*25] DENG C., ZHU D., LI K., GOU C., LI F., WANG Z., ZHONG S., YU W., NIE X., SONG Z., SHI G., FAN H.: Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683* (2025). 2

[FSX*23] FRÜHSTÜCK A., SARAFIANOS N., XU Y., WONKA P., TUNG T.: Vive3d: Viewpoint-independent video editing using 3d-aware gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 4446–4455. 3

[FSZ*25] FANG Y., SUN Z., ZHANG S., WU T., XU Y., ZHANG P., WANG J., WETZSTEIN G., LIN D.: Relightvid: Temporal-consistent diffusion model for video relighting, 2025. URL: https://arxiv.org/abs/2501.16330, arXiv:2501.16330. 5

[FWW*24] FENG R., WENG W., WANG Y., YUAN Y., BAO J., LUO C., CHEN Z., GUO B.: Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 6712–6722. 3

[HCS*25] HU J., CHENG Z., SI C., LI W., GONG S.: Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428* (2025). 2

[HHY*24] HUANG Z., HE Y., YU J., ZHANG F., SI C., JIANG Y., ZHANG Y., WU T., JIN Q., CHANPAISIT N., WANG Y., CHEN X., WANG L., LIN D., QIAO Y., LIU Z.: VBench: Comprehensive benchmark suite for video generative models. 6

[HTXL25] HUA H., TANG Y., XU C., LUO J.: V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2025), vol. 39, pp. 3599–3607. 2

[HYZ*25] HU T., YU Z., ZHOU Z., LIANG S., ZHOU Y., LIN Q., LU Q.: Hunyuancustom: A multimodal-driven architecture for customized video generation, 2025. URL: https://arxiv.org/abs/2505.04512, arXiv:2505.04512. 2, 5, 6, 7

[JHM*25] 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* (2025), pp. 17191–17202. 6, 7

[JP21] JAFARIAN Y., PARK H. S.: Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 12753–12762. 6, 7

[KMT*23] KHACHATRYAN L., MOVSISYAN A., TADEVOSYAN V., HENSCHEL R., WANG Z., NAVASARDYAN S., SHI H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 15954–15964. 3

[KSX*25] KUANG J., SHEN Y., XIE J., LUO H., XU Z., LI R., LI Y., CHENG X., LIN X., HAN Y.: Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Comput. Surv. 57*, 8 (Mar. 2025). URL: https://doi.org/10.1145/3711680, doi:10.1145/3711680. 2

[KTZ*24] 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). 2, 3, 4, 7

[LCLY24] LEE J., CHA S., LEE Y., YANG C.: Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360* (2024). 2

[LDH*21] LIU R., DENG H., HUANG Y., SHI X., LU L., SUN W., WANG X., DAI J., LI H.: Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 14040–14049. 3

[LGL*25] LIANG R., GOJCIC Z., LING H., MUNKBERG J., HASSELGREN J., LIN Z.-H., GAO J., KELLER A., VIJAYKUMAR N., FIDLER S., WANG Z.: Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2025). 5

[LLQ*22] 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* (2022), pp. 17562–17571. 3

[LLSH23] LI J., LI D., SAVARESE S., HOI S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (2023), PMLR, pp. 19730–19742. 2

[LLWL23] LIU H., LI C., WU Q., LEE Y. J.: Visual instruction tuning. *Advances in neural information processing systems 36* (2023), 34892–34916. 2

[LWJ24] LI Y., WANG C., JIA J.: Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision* (2024), Springer, pp. 323–340. 3

[LXRB25] LI X., XUE H., REN P., BO L.: Diffueraser: A diffusion model for video inpainting, 2025. URL: https://arxiv.org/abs/2501.10018, arXiv:2501.10018. 5, 7

[LZL*24a] LIU S., ZHANG Y., LI W., LIN Z., JIA J.: Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 8599–8608. 3

[LZL*24b] LIU Y., ZHANG K., LI Y., YAN Z., GAO C., CHEN R., YUAN Z., HUANG Y., SUN H., GAO J., ET AL.: Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177* (2024). 2, 3

[MRKK24] MAAZ M., RASHEED H., KHAN S., KHAN F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2024), pp. 12585–12602. 3

[NCW*24] NIU M., CUN X., WANG X., ZHANG Y., SHAN Y., ZHENG Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222* (2024). 7

[NXZ*24] NAN K., XIE R., ZHOU P., FAN T., YANG Z., CHEN Z., LI X., YANG J., TAI Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371* (2024). 4, 5

[omn26] Omnivideo2: A flexible framework to bridge video understanding, generation and editing, 2026. URL: https://github.com/SAIS-FUXI/Omni-Video. 4

[Ope25] OPENAI: Gpt-5. https://openai.com/index/introducing-gpt-5/, aug 2025. 2

[OVK] OORD A., VINYALS O., KAVUKCUOGLU K.: Neural discrete representation learning. 5

[PCS*23] PHONGTHAWEE P., CHINCHUTHAKUN W., SINSUNTHITHET N., RAJ A., JAMPANI V., KHUNGURN P., SUWAJANAKORN S.: Diffusionlight: Light probes for free by painting a chrome ball. In *ArXiv* (2023). 5

[PPTM*16] PERAZZI F., PONT-TUSET J., MCWILLIAMS B., VAN GOOL L., GROSS M., SORKINE-HORNUNG A.: A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2016). URL: http://dx.doi.org/10.1109/cvpr.2016.85, doi:10.1109/cvpr.2016.85. 5, 7

[PX23] PEEBLES W., XIE S.: Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 4195–4205. 3

[QCZ*23] 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* (2023), pp. 15932–15942. 3

[QDZ*24] QIAN R., DONG X., ZHANG P., ZANG Y., DING S., LIN D., WANG J.: Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems 37* (2024), 119336–119360. 3

[RSR*23] RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W., LIU P. J.: Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL: https://arxiv.org/abs/1910.10683, arXiv:1910.10683. 6

[SCW*24] SONG E., CHAI W., WANG G., ZHANG Y., ZHOU H., WU F., CHI H., GUO X., YE T., ZHANG Y., ET AL.: Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 18221–18232. 3

[TAB*23] 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). 2

[TBX*25] TANG Y., BI J., XU S., SONG L., LIANG S., WANG T., ZHANG D., AN J., LIN J., ZHU R., ET AL.: Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* (2025). 2

[TYQ*25] TAN Z., YANG H., QIN L., GONG J., YANG M., LI H.: Omni-video: Democratizing unified video understanding and generation. *arXiv preprint arXiv:2507.06119* (2025). 2, 6, 7

[VDOV*17] VAN DEN OORD A., VINYALS O., ET AL.: Neural discrete representation learning. *Advances in neural information processing systems 30* (2017). 2

[WGW*23] 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* (2023), pp. 7623–7633. 3

[WHH*24] WENG Y., HAN M., HE H., CHANG X., ZHUANG B.: Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision* (2024), Springer, pp. 453–470. 3

[WLY*25] WEI C., LIU Q., YE Z., WANG Q., WANG X., WAN P., GAI K., CHEN W.: Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377* (2025). 2, 4

[WPQ*24] WANG J., PU J., QI Z., GUO J., MA Y., HUANG N., CHEN Y., LI X., SHAN Y.: Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746* (2024). 7

[WWBA25] WAN T., WANG A., BAOLE AI E.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025). 6

[WYM*22] WANG P., YANG A., MEN R., LIN J., BAI S., LI Z., MA J., ZHOU C., ZHOU J., YANG H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning* (2022), PMLR, pp. 23318–23340. 2

[WZG*25] WANG X., ZHANG S., GAO C., WANG J., ZHOU X., ZHANG Y., YAN L., SANG N.: Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences 68*, 10 (2025), 1–14. 3

[WZY*25] WU C., ZHENG P., YAN R., XIAO S., LUO X., WANG Y., LI W., JIANG X., LIU Y., ZHOU J., LIU Z., XIA Z., LI C., DENG H., WANG J., LUO K., ZHANG B., LIAN D., WANG X., WANG Z., HUANG T., LIU Z.: Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871* (2025). 2

[XAH22] XU Y., ALBAHAR B., HUANG J.-B.: Temporally consistent semantic video editing. In *European Conference on Computer Vision* (2022), Springer, pp. 357–374. 3

[XLZL19] XU R., LI X., ZHOU B., LOY C. C.: Deep flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 3723–3732. 3

[XYF*18] XU N., YANG L., FAN Y., YUE D., LIANG Y., YANG J., HUANG T.: Youtube-vos: A large-scale video object segmentation benchmark, 2018. URL: https://arxiv.org/abs/1809.03327, arXiv:1809.03327. 5, 7

[YFZ*24] YIN S., FU C., ZHAO S., LI K., SUN X., XU T., CHEN E.: A survey on multimodal large language models. *National Science Review 11*, 12 (2024), nwae403. 2

[YTZ*24] 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). 3, 7

[YZLL23] 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* (2023), pp. 1–11. 3

[ZAOT18] ZHOU B., ANDONIAN A., OLIVA A., TORRALBA A.: Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 803–818. 2

[ZLB23] ZHANG H., LI X., BING L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023). 2

[ZLCL23] ZHOU S., LI C., CHAN K. C., LOY C. C.: ProPainter:

Improving propagation and transformer for video inpainting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)* (2023). 3, 5, 7

[ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (2023), pp. 3836–3847. 3

[ZRA25] ZHANG L., RAO A., AGRAWALA M.: Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations* (2025). URL: https://openreview.net/forum?id=u1cQYxRI1H. 5

[ZSH*22] ZHANG Y., SUN J., HE X., FU H., JIA R., ZHOU X.: Modeling indirect illumination for inverse rendering. In *CVPR* (2022). 5

[ZSZS19] ZABLOTSKAIA P., SIAROHIN A., ZHAO B., SIGAL L.: Dwnet: Dense warp-based network for pose-guided human video generation. *British Machine Vision Conference,British Machine Vision Conference* (Oct 2019). 6, 7

[ZYD*24] ZHANG D., YU Y., DONG J., LI C., SU D., CHU C., YU D.: Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024* (2024), pp. 12401–12430. 2