LiveEdit: Towards Real-Time Diffusion-Based Streaming Video Editing

Xinyu Wang[1], Chongbo Zhao[1], Fangneng Zhan[2], and Yue Ma[2⋆]

[1] Tsinghua University, Beijing, China
cpwxyxwcp@gmail.com
[2] The Hong Kong University of Science and Technology, Hong Kong SAR, China

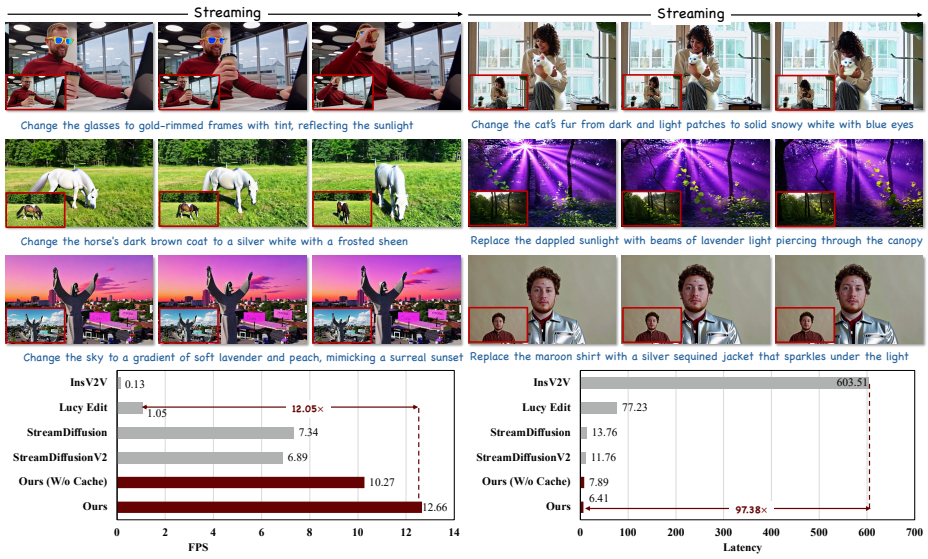

Fig. 1: Gallery of various editing results and efficiency comparisons. We propose the *LiveEdit*, a novel streaming video editing framework capable of performing causal, chunk-by-chunk manipulation with ultra-low latency and strict background preservation. By synergizing a progressive three-stage architectural distillation pipeline with an AR-oriented Mask Cache, *LiveEdit* effectively resolves the architectural incompatibilities and computational bottlenecks inherent in streaming editing paradigms.

Abstract. Streaming video editing has made rapid progress, yet practical deployment is still limited by two core issues: maintaining stable backgrounds and non-edited regions over time, and achieving the low latency required for real-time interactive scenarios. Meanwhile, recent streaming video generation methods are mostly developed for synthesis and cannot be directly applied to editing due to the strict preservation requirement

⋆ Corresponding author.

and region-specific control. In this work, we present a novel streaming video editing framework that performs causal, frame-by-frame editing with strong content preservation and real-time responsiveness. Our key design is a three-stage distillation pipeline that progressively transfers editing capability from a powerful bidirectional foundation model to an efficient unidirectional streaming editor, enabling stable long-horizon edits without sacrificing visual fidelity. To further support real-time deployment, we introduce an AR-oriented mask cache that reuses region-related computation across frames, substantially reducing redundant processing and accelerating inference. Finally, we establish a dedicated benchmark for streaming video editing. Extensive evaluations demonstrate that our method achieves state-of-the-art visual quality among streaming baselines while drastically boosting inference speed to 12.66 FPS, making it suitable for interactive and augmented reality applications.

Keywords: Streaming Video Editing · Diffusion Models · Efficient Generation

1 Introduction

Video editing [7,12,17,20,31,39,45,48,49,52,54,59] has witnessed significant advancements, driven by the increasing demand for high-quality content creation and interactive digital experiences. As augmented reality and live-streaming applications become more prevalent, the industry's focus is shifting from traditional offline batch processing toward real-time, responsive editing. However, achieving practical, low-latency deployment remains a formidable challenge, particularly when moving toward a streaming paradigm where video must be processed chunk-by-chunk without access to future information.

As illustrated in Fig. 2, the transition to practical streaming video editing is currently obstructed by two fundamental bottlenecks. **i) Attention distribution shift**: State-of-the-art video diffusion models typically rely on bidirectional or global attention to maintain temporal consistency. Directly adapting these non-causal models to a causal streaming setting—where future frames are unavailable—often leads to a "forgetting" effect or severe flickering, as the model lacks the global structural context required for stable editing. **ii) Spatial-temporal token redundancy**: Standard diffusion pipelines treat every frame as an independent, heavy generation task. However, in autoregressive (AR) oriented streaming generation, the majority of the background remains static or undergoes predictable linear motion. Repeatedly applying dense Feed-Forward Network (FFN) and Attention modules to these redundant, unedited regions leads to prohibitive per-frame latency, making real-time interactive experiences unattainable on edge devices.

To bridge this gap, we introduce a novel framework designed for causal, chunk-by-chunk video editing with strong content preservation and ultra-low latency. Our approach addresses the latency-stability trade-off through a structured three-stage distillation pipeline that progressively transfers the editing

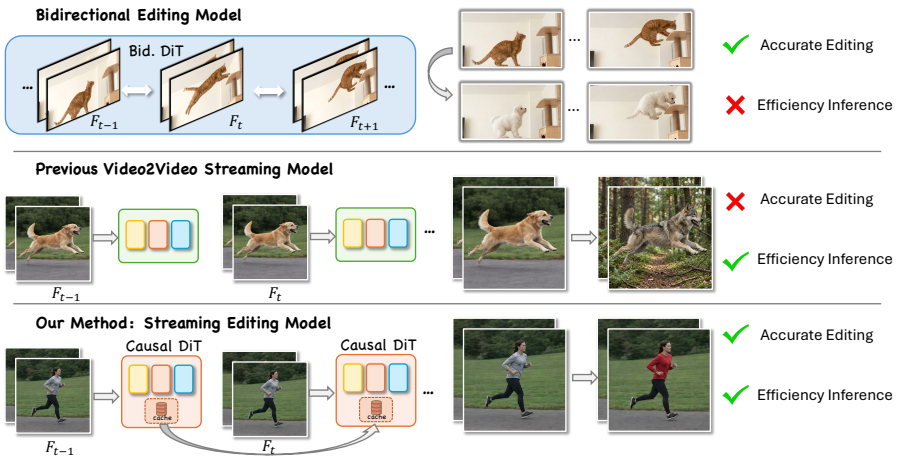

Fig. 2: Comparison of video editing paradigms. Unlike bidirectional models that suffer from inefficient inference, and past streaming models that fail to preserve accurate unedited content, our proposed streaming editing model leverages a Causal DiT with a mask-guided cache mechanism to achieve high-fidelity and efficient editing.

capabilities of a powerful bidirectional foundation model to an efficient, unidirectional streaming editor.

Specifically, **Stage 1 (Foundation Tuning)** focuses on equipping a Bidirectional Diffusion Transformer with robust editing abilities. By leveraging full attention mechanisms and text embeddings, the model learns complex editing mappings supervised by an $\mathcal{L}_{MSE}$ loss. To bridge the gap between offline and streaming processing, **Stage 2 (Teacher Forcing for Chunk-wise Causal Initial)** transitions the architecture from a bidirectional to a Causal DiT. We employ a teacher-forcing strategy equipped with chunk-wise causal attention, ensuring the model successfully adapts to sequential, unidirectional inputs while strictly maintaining the visual quality and editing priors established in the first stage. Finally, **Stage 3 (DMD for Streaming Video Editing)** performs advanced distillation to achieve real-time inference. By integrating Distribution Matching Distillation (DMD) with a frozen Real Score and a trainable Fake Score model, we compress the generation process into merely 4 steps. This stage is optimized via both $\mathcal{L}_{MSE}$ and $\nabla_\theta \mathcal{L}_{DMD}$ gradients, utilizing pruned noise inputs to further accelerate streaming video editing.

Beyond architectural distillation, redundant background processing remains a critical bottleneck for AR applications. To address this, we introduce an **AR-oriented Mask Cache** during inference. Instead of running the full DiT forward pass for every frame, our method calculates the L_2 distance between the edited output and the source to extract an accurate editing mask. We observe a functional divergence in how different modules handle redundant information: while the Feed-Forward Network and Cross-Attention are essential for maintaining per-pixel spatial detail and text-conditioned alignment, the Self-Attention layers

exhibit significant spatio-temporal redundancy in unedited regions. This selective spatial-temporal reuse significantly reduces per-frame latency while guaranteeing that the editing quality remains indistinguishable from the full-calculation baseline.

Our contributions are summarized as follows:

- We analyze the property of streaming video editing and present the novel streaming video editing framework capable of performing causal, chunk-by-chunk editing with high fidelity and ultra-low inference latency(12.66 FPS).
- Technically, we first design a comprehensive three-stage distillation pipeline that effectively migrates complex editing knowledge from a Bidirectional DiT teacher to an ultra-efficient, 4-step Causal DiT student. Moreover, the AR-oriented Mask Cache mechanism is proposed by leveraging L_2 distance to dynamically decouple computation.
- We establish a dedicated benchmark for streaming video editing and demonstrate that our method achieves state-of-the-art performance in terms of both visual quality, temporal consistency, and throughput.

2 Related Work

Video Generation and Controllable Editing. Diffusion models have significantly advanced video generation [3, 33–37, 43, 55] and editing. To achieve versatile control [32], unified frameworks like VACE [17] propose all-in-one architectures, while EditVerse [18] and UNIC [56] employ in-context learning to handle diverse tasks. These approaches typically integrate source videos and dense multi-modal conditions into massive joint representations or extended context sequences. Meanwhile, models like InsV2V [9] and Lucy Edit [45] leverage synthetic datasets or channel-wise concatenation for high-fidelity modifications [38]. Additionally, reinforcement learning has been explored to further align generated content with human intents [2].

Despite their impressive visual quality, these methods inherently rely on offline, non-causal processing. Processing entire video frames and complex conditions jointly quadratically increases the computational overhead. Consequently, these models must observe the entire temporal context before outputting the initial frame. This unacceptable latency intrinsically hinders their deployment in real-time augmented reality scenarios. To break this paradigm, we propose a novel streaming video editing framework explicitly designed for extreme real-time responsiveness.

Streaming Autoregressive Video Models. Streaming autoregressive generation effectively overcomes the latency bottlenecks of offline models. Foundational works like Diffusion Forcing [5] and StreamDiffusion [19] redefine denoising as block-wise sequential processing. This paradigm enables massive-scale world models (MAGI-1 [46]) and infinite-length cinematic generation (SkyReels-V2 [6], StreamingT2V [14]). To enhance stability and efficiency, Rolling Forcing [28]

suppresses error accumulation, Stable Video Infinity [22] mitigates autoregressive drift for infinite-length generation via error-recycling fine-tuning, Self Forcing [15] bridges exposure bias via self-generated conditioning, and StreamDiffusionV2 [13] introduces sink-token-guided rolling KV caches for ultra-low-latency live generation. Furthermore, EgoEdit [20] pioneered streaming models for egocentric video editing by maintaining temporal consistency in continuous first-person visual translations.

However, directly migrating these generation-tailored mechanisms to general video editing introduces a fundamental misalignment. Video generation synthesizes motion from scratch, heavily relying on past generated predictions (e.g., Self Forcing's serial feedback). Conversely, streaming editing is strongly conditioned on continuous source video streams, prioritizing precise spatial restoration over free-form motion synthesis. Serially depending on historical outputs or employing generation-oriented caches (e.g., DeepCache [30], VMem [21], or StreamDiffusionV2 [13]) degrades high-frequency structural details when rigorously aligning with semantic editing trajectories. Discarding redundant serial feedback, we introduce an AR-based Cache that achieves extreme computational compression while ensuring strict temporal consistency and spatial alignment.

Efficiency and Distillation in Diffusion. Accelerating reverse diffusion [27, 41, 44, 51, 60, 61] is crucial for real-time performance. Early one-step distillation efforts focused on the image domain, utilizing techniques like Rectified Flow (InstaFlow [29]), target score distillation (TSD-SR [11]), and Distribution Matching Distillation [57, 58], alongside progressive adversarial distillation (SDXL-Lightning [25]). These advancements have naturally extended to the video domain, enabling efficient one-step generation (AAPT [26]), restoration (SeedVR2 [50]), and super-resolution (DOVE [8]). Sparse VideoGen2 [53] further reduces redundancy via attention permutation. Orthogonal to distillation, structural optimizations like Token Merging [4] and FlashAttention [10] are widely adopted to alleviate system memory overheads.

Most closely related to our work is FlashVSR [62], which achieves real-time streaming video super-resolution via a three-stage distillation pipeline designed to progressively accelerate latent rendering. Similarly, PersonaLive [23] leverages appearance distillation for live portrait animation. However, these tasks fundamentally involve low-level pixel mapping or structurally constrained regions. Directly transferring these distillation schemes to general video editing—which involves complex semantic reconstruction, cross-scale feature evolution, and training-free reward-guided control—frequently degrades high-fidelity details. In contrast, we propose a tailored three-stage distillation strategy integrated with our AR-based Cache, achieving unprecedented inference acceleration without sacrificing complex editing fidelity.

3 Method

3.1 Motivation

We summarize the two primary observations regarding the inefficiencies of adapting state-of-the-art video diffusion models to the streaming video editing task and propose the modules to address them.

Attention distribution shift. In the offline video editing phase, state-of-the-art bidirectional diffusion models utilize dense temporal attention to propagate structural information, natively assigning significant weights to both past and immediate future tokens. However, we observe that abruptly truncating these future keys and values for causal execution causes a severe shift in the attention distribution. Specifically, as illustrated in Fig. 3, the loss of future context forces the attention weights to flatten and distribute uniformly across all available historical frames. This homogenized attention behavior intrinsically conflicts with the streaming paradigm, which relies heavily on the nearest neighboring frames to maintain temporal coherence and structural integrity. Consequently, this over-smoothed dependency disrupts the pre-trained structural priors, making it theoretically suboptimal to directly deploy naively truncated bidirectional models in a streaming pipeline. To address this, we introduce a progressive three-stage distillation pipeline. Our teacher-forcing mechanism explicitly aligns the causal attention distribution with the localized bidirectional prior, ensuring robust, localized representation mapping without the need for future context.

Spatial-temporal token redundancy. Existing streaming generation approaches (e.g., StreamV2V [24], StreamDiffusion [19] and StreamDiffusionV2 [13]) primarily focus on global Video-to-Video translation tasks. Specifically, for every incoming frame, these pipelines perform dense computations across all spatial tokens globally to synthesize the entire scene. However, we note that streaming video editing fundamentally differs from global translation, as the intermediate feature representations of tokens in unedited regions must strictly maintain absolute temporal consistency. Therefore, applying such a global generation paradigm to editing tasks inherently disrupts the visual integrity of unedited areas, causing destructive structural degradation and severe flickering in the background, alongside massive computational redun-

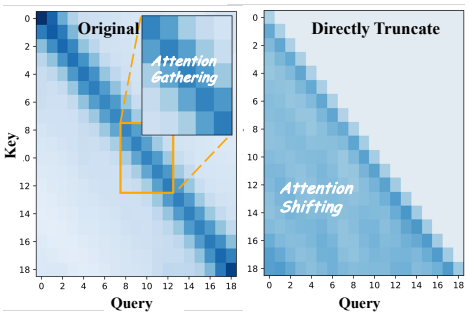

Fig. 3: Visualization of the attention distribution shift. Left: The bidirectional prior exhibits localized attention gathering. **Right:** Direct causal truncation forces attention to spread uniformly across all historical frames.

dancy. To address this, we introduce the AR-oriented Mask Cache mechanism. Only the actively edited spatial tokens undergo full calculation, while the intermediate features of highly correlated static tokens are directly retrieved from the cache, ensuring strict background preservation and efficient generation.

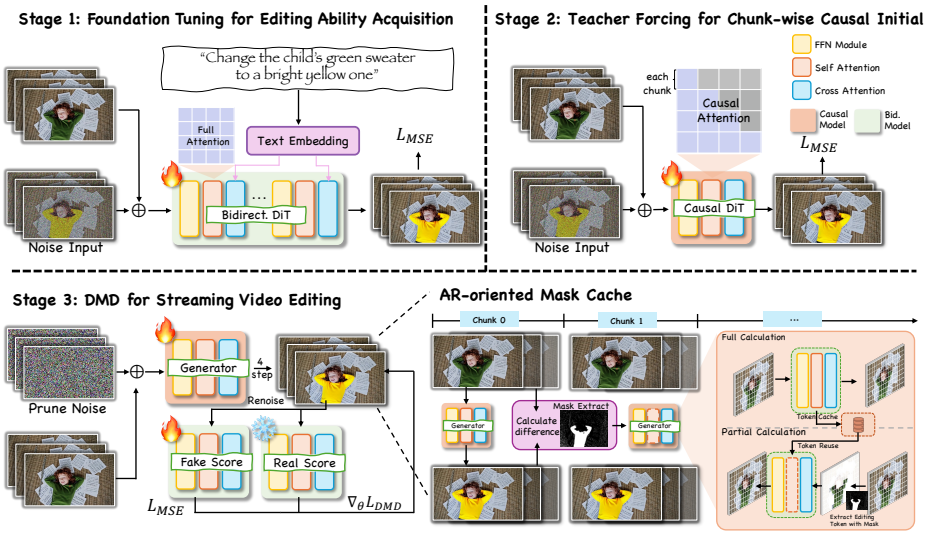

Fig. 4: Overview of the proposed streaming video editing framework. Our approach features a three-stage distillation pipeline that transfers editing capabilities from a bidirectional DiT to a 4-step causal model. Furthermore, an AR-oriented Mask Cache accelerates real-time inference by dynamically decoupling computation and reusing tokens in unedited background regions.

3.2 Three-Stage Distillation Pipeline

To bridge the architectural gap between offline bidirectional priors and online causal execution, we propose a progressive three-stage distillation pipeline. Let $z_0 \in \mathbb{R}^{F\times C\times H\times W}$ denote the input video latent sequence and c represent the text embedding extracted from the editing prompt. This design effectively transfers the high-fidelity editing capabilities of a foundation model into an ultra-fast, unidirectional streaming editor. An overview of the framework is illustrated in Fig. 4.

Stage 1: Foundation Tuning for Editing Ability Acquisition. In the initial stage, we establish a robust multimodal video-to-video editing baseline utilizing a Bidirectional Diffusion Transformer (DiT). Specifically, the network processes the channel-wise concatenation of the original video latent and the noisy latent z_t at timestep t. To mitigate the quadratic computational overhead inherently

associated with long token sequences in video generation, we emphasize this channel-wise integration strategy over spatial or temporal sequence concatenation. The bidirectional architecture, denoted as ϵ_θ^{bid}, leverages full temporal and spatial attention to learn complex mapping functions for content manipulation. The entire foundation model is supervised via the standard noise-matching objective:

$$\mathcal{L}_{MSE}^{bid} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,I),t,c}\left[\left\|\epsilon - \epsilon_\theta^{bid}(z_t,t,c)\right\|_2^2\right]$$

This optimization yields a powerful offline editing prior capable of maintaining high-fidelity generation.

Stage 2: Teacher Forcing for Chunk-wise Causal Initial. To enable basic streaming input-output functionality, the architecture must transition from a bidirectional to an autoregressive (AR) generation paradigm. However, directly applying causal masks to a pre-trained bidirectional model severely degrades performance. Therefore, we introduce a Teacher Forcing mechanism equipped with chunk-wise causal attention. Let M_{causal} denote the causal attention mask that restricts temporal tokens from attending to future chunks. The causal DiT, ϵ_θ^{causal}, is optimized to predict the noise while strictly adhering to the causal constraint:

$$\mathcal{L}_{MSE}^{causal} = \mathbb{E}_{z_0,\epsilon,t,c}\left[\left\|\epsilon - \epsilon_\theta^{causal}(z_t,t,c \mid M_{causal})\right\|_2^2\right]$$

Inspired by recent autoregressive frameworks such as Causal Forcing, we establish that deriving an AR model via explicit Teacher Forcing is structurally essential for streaming video generation. By aligning the output distribution of the causal DiT with the bidirectional representations learned in Stage 1, the model prevents the structural collapse typically caused by the absence of future tokens.

Stage 3: DMD for Streaming Video Editing. To achieve ultra-low latency while eliminating accumulated shift errors during continuous streaming, we perform advanced step-distillation using Distribution Matching Distillation (DMD). Recent autoregressive generation paradigms, notably Self-Forcing, heavily rely on an Ordinary Differential Equation (ODE) initialization phase to bridge the training-inference distribution gap. However, this ODE initialization incurs prohibitive computational resource overhead and severely limits practical scalability. To circumvent this fundamental bottleneck, we directly utilize the AR-based model parameters from Stage 2 (ϵ_θ^{causal}) to initialize the 4-step DMD generator G_θ. This architectural decision not only avoids the excessive initialization overhead but also provides a highly stable starting point for distillation, echoing the empirical insights observed in recent works like EgoEdit.

During training, the generator G_θ maps pruned noise inputs directly to the edited frames. The distillation process is jointly optimized by $\mathcal{L}_{MSE}$ and the DMD gradient $\nabla_\theta\mathcal{L}_{DMD}$. The DMD gradient is computed between a frozen Real Score model ϵ_ϕ^{real} and a trainable Fake Score model ϵ_ψ^{fake}:

$$\nabla_\theta\mathcal{L}_{DMD} = \mathbb{E}_{z_T,c}\left[w(t)\left(\epsilon_\phi^{real}(z_t,t,c) - \epsilon_\psi^{fake}(z_t,t,c)\right)\nabla_\theta G_\theta(z_T,c)\right]$$

where z_t is the intermediate latent simulated from the generated output $G_\theta(z_T, c)$, and $w(t)$ is a timestep-dependent weighting function. This mechanism effectively compresses the generation process into merely 4 inference steps, ensuring real-time responsiveness.

3.3 AR-oriented Mask Cache

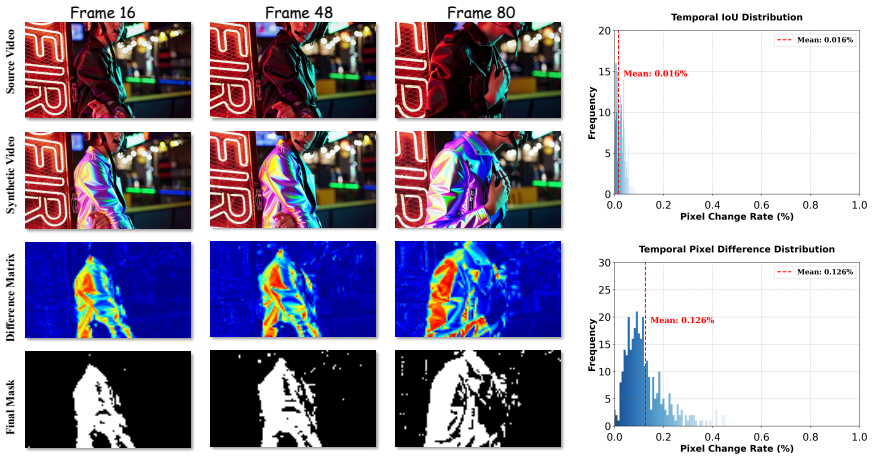

Fig. 5: Visualization of the temporal consistency analysis and mask generation process. The left panels show (from top to bottom) the source video frames, the synthesized video frames, the computed difference matrices, and the resulting binary masks. The right panels display the statistical distributions of Temporal IoU and Pixel Difference across the sequence, with mean values of 0.016% and 0.126%, respectively, indicating high structural stability.

While the three-stage distillation pipeline resolves the architectural incompatibility, the fundamental issue of spatial-temporal redundancy remains a critical bottleneck for real-time inference. To address this, we introduce the AR-oriented Mask Cache mechanism, dynamically decoupling the computational graph based on regional editing activity.

During the streaming inference phase, the pipeline processes the video in sequential chunks. To dynamically route the computation for an incoming chunk k, we derive its spatial editing mask from the generation trajectory of the preceding chunk. Let z_{src}^{k-1} denote the original source latent representation of the previously generated chunk $k-1$, and z_{edit}^{k-1} denote its corresponding edited output latent. We extract a binary spatial editing mask $M^k \in \{0,1\}^{H\times W}$ for the current chunk k by computing the L_2 distance between these two latents:

$$M_{u,v}^k = \mathbb{I}\left(\left\| z_{edit,u,v}^{k-1} - z_{src,u,v}^{k-1} \right\|_2 > \tau\right)$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is dynamically determined by calculating a redundancy level among the tokens. This mask geometrically separates the spatial layout of the incoming chunk k into active editing regions ($M^k_{u,v} = 1$) and static background regions ($M^k_{u,v} = 0$). As visualized in Fig. 5, our mask generation process demonstrates extremely high structural stability, with the statistical distribution across the entire sequence.

Instead of executing the complete network forward pass globally for chunk k, our mechanism implements a spatial routing strategy guided by M^k. For the spatial tokens located within the active mask, the model allocates its full computational capacity, executing the complete sequence of Self-Attention, Cross-Attention, and Feed-Forward Network modules (*Full Calculation*). Conversely, for the tokens corresponding to the unedited regions, the mechanism entirely bypasses the computationally expensive layers. Let $\mathcal{F}$ denote the full block transformation and $z^k_{u,v}$ denote the input token for the current chunk; the output token feature $f^k_{u,v}$ is determined by:

$$f^k_{u,v} = \begin{cases} \mathcal{F}(z^k_{u,v}) & \text{if } M^k_{u,v} = 1 \\ f^{k-1}_{u,v} & \text{if } M^k_{u,v} = 0 \end{cases}$$

The intermediate feature representations for the static regions are directly retrieved from a maintained Token Cache populated by the preceding chunk (*Token Reuse*). This dynamic, inter-chunk spatial decoupling dramatically reduces the per-frame computational complexity, achieving substantial acceleration while strictly guaranteeing absolute visual consistency in the unedited background areas.

4 Experiment

4.1 Implementation Details

Model Configuration. We build our foundation model upon Wan2.1-T2V-1.3B [47]. During Stage 1, we explicitly employ channel-wise concatenation to integrate the noisy latent z_t and the condition latent, rather than appending them token-wise along the sequence dimension. By strictly maintaining the original sequence length, this structural design fundamentally bypasses the quadratic computational explosion in attention mechanisms typically associated with spatial-temporal video generation. For training, our dataset comprises 20K high-quality video-video pairs, which are carefully filtered from the large-scale Ditto-1M [1] dataset.

Training Setup. The three-stage distillation pipeline is trained progressively on 8 NVIDIA A100 GPUs. We employ the AdamW optimizer across all stages.

- **Stage 1 (Foundation Tuning):** The bidirectional foundation model ϵ^{bid}_θ is trained for 9K steps with a learning rate of 10^{-5} and a global batch size of 8. We utilize a standard noise scheduler with continuous timesteps $t \in [0, 1000]$.

Fig. 6: Qualitative comparison of streaming video editing performance. The source videos and instructions are displayed at the top. While existing methods exhibit significant limitations, leading to structural collapse or an inability to accurately follow the text, our approach precisely modifies the target regions and preserves the visual quality and temporal coherence of the original scenes.

- **Stage 2 (Teacher Forcing):** We transition the bid. Dit to the causal DiT ϵ_θ^{causal} by introducing the chunk-wise causal attention mask M_{causal}. The model is fine-tuned for an additional 20K steps. The temporal chunk size is set to 3 latent frames, ensuring that tokens can only attend to the current and strictly preceding chunks.
- **Stage 3 (DMD):** The 4-step generator G_θ is initialized directly from the Stage 2 weights. deliberately bypassing the computationally expensive ODE initialization. This bypass is explicitly enabled by the autoregressive training conducted in Stage 2, which provides a well-aligned causal distribution starting point. For the 4-step generation, the sampling timesteps are specifically set to $[0, 250, 500, 750]$. Both the Real Score model and the Fake Score model are initialized from the foundation model weights obtained in Stage 1. We apply a reduced learning rate of 10^{-5} for 10K steps, utilizing the timestep-dependent weighting function $w(t)$ as proposed in standard DMD formulations.

Inference and Mask Cache. During the streaming inference phase, the model operates in a purely autoregressive manner, decoding 3 frames per step. For the AR-oriented Mask Cache, rather than using a fixed empirical value, the L_2 distance threshold τ for the binary spatial editing mask M^k is dynamically calculated to explicitly prune 70% of the redundant spatial tokens. Specifically, this caching mechanism is applied within the Self-Attention layers, as we empirically find that applying the cache to the Self-Attention computation introduces no

degradation in generation quality. Tokens in these pruned regions bypass the standard attention computation, and their intermediate features are directly retrieved from the Token Cache. This dynamic routing allows the 4-step streaming editing to achieve an ultra-low latency of $79ms$ per-frame, ensuring real-time responsiveness without background flickering.

4.2 Qualitative comparison.

We present the visual comparison between our proposed framework and two distinct categories of state-of-the-art video editing baselines: bidirectional offline editing models (LucyEdit [45], InsV2V [9], VideoCoF [54]) and streaming generation models (StreamV2V [24], StreamDiffusion [19], StreamDiffusionV2 [13]). As illustrated in Fig. 6, existing methods exhibit significant limitations in balancing precise prompt alignment with structural preservation.

StreamV2V fundamentally struggles to execute localized, precise attribute modifications, failing to apply the requested edits and leaving the source attributes largely unaltered. Conversely, approaches such as InsV2V and VideoCoF suffer from severe color bleeding because they indiscriminately apply target attributes across the entire frame, thereby corrupting the subject's skin tone and the surrounding environment. Furthermore, methods relying on global translation paradigms inherently compromise the visual integrity of unedited areas; StreamDiffusion and StreamDiffusionV2 exhibit severe structural collapse and unintended style shifts, whereas LucyEdit lacks the strict spatial control necessary to preserve delicate structural details. Overcoming these limitations, our method accurately interprets complex textual conditions to apply distinct textures and colors exclusively to target regions. Benefiting from the proposed AR-oriented Mask Cache mechanism, our approach explicitly decouples the generation process, guaranteeing zero visual degradation in unedited background regions. Consequently, complex lighting conditions, shadows, and subject identities are strictly preserved alongside high-fidelity edits.

4.3 Quantitative comparison.

We compare our method against the state-of-the-art baselines on a collected benchmark consisting of 120 pairs. To comprehensively assess both the generation quality and editing accuracy, we employ six standard automated metrics, with the results summarized in Tab. 1.

For Text Alignment, following EgoEdit [20], we compute the CLIP [40] feature similarity of the edited results with the text prompt to evaluate editing consistency and instruction adherence. To assess the aesthetic appeal, we utilize the LAION-Aesthetic Score Predictor [42] for the Aesthetic Quality metric. Furthermore, we employ the VBench evaluation framework [16] to measure Background Consistency, Motion Smoothness, Dynamic Degree, and Imaging Quality, which collectively provide a comprehensive assessment of the overall visual fidelity.

Table 1: Quantitative comparison of video editing methods. We evaluate across six metrics: Text Alignment (TA), Background Consistency (BC), Motion Smoothness (MS), Dynamic Degree (DD), Aesthetic Quality (AQ), and Imaging Quality (IQ). **Red** and **Blue** denote the best and second-best results, respectively.

Method	TA↑	BC↑	MS↑	DD↑	AQ↑	IQ↑
LucyEdit	0.253	0.943	0.990	0.266	0.529	0.707
VideoCoF	0.245	0.953	0.991	0.094	0.542	0.709
InsV2V	0.259	0.943	0.986	0.196	0.577	0.708
StreamDiffusion	0.239	0.886	0.975	0.239	**0.590**	**0.717**
StreamDiffusionV2	0.252	0.951	**0.992**	0.264	0.539	0.653
StreamV2V	0.244	0.934	0.989	0.153	0.548	0.712
Ours (W/o Cache)	**0.265**	**0.956**	0.991	**0.282**	**0.584**	**0.720**
Ours (W/ Cache)	**0.270**	**0.956**	**0.992**	**0.256**	0.581	0.708

Table 2: Ablation of our three-stage distillation pipeline. Comparison of generation configurations and inference efficiency across the different training stages.

	Stage 1 (Foundation)	**Stage 2 (Teacher Forcing)**	**Stage 3 (DMD)**
Is streaming?	×	✓	✓
NFEs	100	100	4
With CFG?	✓	✓	×
Latency	197.48	200.36	7.89
First chunk size	Full sequence	3 frames	3 frames
Next chunk size	N/A	3 frames	3 frames
Text Alignment	0.268	0.264	0.265
Image Quality	0.716	0.702	0.720

As demonstrated in Tab. 1, our proposed framework achieves best performance across almost all dimensions. Notably, while bidirectional offline architectures naturally benefit from future temporal context, our unidirectional streaming approach not only bridges the performance gap but strictly outperforms them in Text Alignment (achieving 0.270 compared to InsV2V's 0.259). Furthermore, our method achieves the highest Dynamic Degree and Imaging Quality among all evaluated methods, while maintaining highly competitive Aesthetic Quality against strong streaming baselines like StreamDiffusion. Crucially, the results demonstrate that the proposed AR-oriented cache improves Text Alignment and Motion Smoothness while perfectly preserving Background Consistency, proving that our mechanism effectively maintains high visual integrity during continuous sequential processing. Furthermore, we invited 20 volunteers to rank methods across background consistency, editing fidelity, and overall quality. The results are provided in the supplementary materials.

4.4 Ablation Study

We conduct an ablation study to validate the necessity of each phase within our three-stage distillation pipeline, with the configuration progression summarized in Tab. 2.

Table 3: Quantitative results of the cache-mechanism ablation. Comparing the performance of applying the caching mechanism to either Self-Attention or FFN layers. **Red** denote the best results.

Method	TA↑	BC↑	MS↑	DD↑	AQ↑	IQ↑
W/o Cache	0.265	**0.956**	0.991	**0.282**	**0.584**	**0.720**
Cache on SA	**0.270**	0.956	**0.992**	0.256	0.581	0.708
Cache on FFN	0.236	0.841	0.982	0.017	0.440	0.513

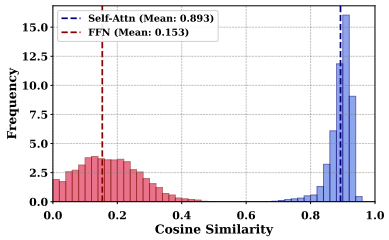

Fig. 7: Distribution of token cosine similarity between consecutive denoising step.

Effectiveness of Stage 1 (Foundation Tuning). The initial phase establishes a robust bidirectional prior essential for high-fidelity structural preservation and text alignment. Operating globally across the full sequence, this stage successfully acquires complex editing capabilities. However, it requires 100 Network Function Evaluations (NFEs) and relies on Classifier-Free Guidance (CFG), which strictly prevents online, continuous streaming inference.

Effectiveness of Stage 2 (Teacher Forcing). To transition the offline foundation model to a streaming paradigm, Stage 2 introduces chunk-wise causal attention. This structural modification successfully enables autoregressive execution, processing the video continuously in sequential chunks of 3 frames. While this stage achieves the fundamental streaming input-output format, it still requires 100 NFEs and maintains the CFG dependency, resulting in substantial computational latency that prohibits real-time deployment.

Effectiveness of Stage 3 (DMD). The final distillation phase dramatically accelerates the inference speed. By compressing the generation process down to 4 NFEs and explicitly eliminating the need for CFG (which otherwise doubles the required forward passes), the computational overhead is drastically reduced. Furthermore, initializing the student generator directly from the autoregressive weights optimized in Stage 2 bypasses the expensive standard ODE initialization. Ultimately, the integration of all three stages yields a framework that simultaneously achieves an ultra-low latency of 7.89s for 81 frames, pure streaming functionality, and high-quality editing performance.

Effectiveness of AR-oriented Mask Cache. To identify the optimal integration point for the AR-oriented Mask Cache, we investigate the impact of applying the caching mechanism to different architectural components, specifically comparing its placement in the Self-Attention (SA) layers versus the FFN modules.

As summarized in Tab. 3, applying the cache to the SA layers (Ours) achieves superior performance across all evaluated dimensions. In contrast, caching FFN features causes significant degradation, suggesting that FFN layers contain high-frequency spatial information too sensitive for direct temporal reuse. This architectural divergence is explicitly validated by the feature similarity distributions in Fig. 7, where SA tokens maintain exceptionally high temporal redundancy across consecutive steps, whereas FFN representations exhibit notably low similarity.

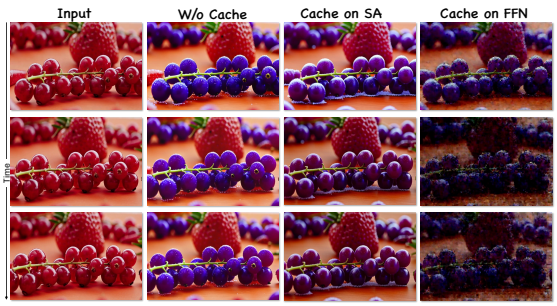

Fig. 8: Visual comparison of different cache locations. With the instruction *"Change the red currants to deep purple grapes with a thin layer of frost on their skins"*.

The visual evidence in Fig. 8 further corroborates these findings. Both the baseline configuration (W/o Cache) and the SA-cache version successfully modify the target object with high fidelity, correctly rendering fine-grained textures and maintaining natural color saturation. Conversely, caching on FFN results in severe blurring and structural instability. These results confirm that caching SA features is the optimal strategy, as it effectively leverages spatial-temporal redundancy to achieve significant acceleration without sacrificing the generative capacity and high-quality spatial details of the full-calculation baseline.

5 Conclusion

In this paper, we presented LiveEdit, a novel streaming video editing framework that successfully adapts powerful bidirectional diffusion priors to an efficient, unidirectional autoregressive paradigm. To overcome the attention distribution shift inherent in causal execution, we introduced a progressive three-stage distillation pipeline comprising Foundation Tuning, Teacher Forcing, and DMD. This architecture effectively bridges the gap between offline global processing and online continuous video editing, compressing the inference process to merely 4 steps. Furthermore, to alleviate spatial-temporal token redundancy, we proposed an AR-oriented Mask Cache mechanism. It reduces computational overhead while strictly guaranteeing zero visual degradation in unedited regions. Extensive quantitative and qualitative evaluations demonstrate that our framework achieves SOTA performance. It uniquely balances high-fidelity text alignment, robust structural preservation, and ultra-low latency, paving the way for highly practical, real-time streaming video editing applications.

References

1. Bai, Q., Wang, Q., Ouyang, H., Yu, Y., Wang, H., Wang, W., Cheng, K.L., Ma, S., Zeng, Y., Liu, Z., Xu, Y., Shen, Y., Chen, Q.: Scaling instruction-based video editing with a high-quality synthetic dataset. arXiv preprint arXiv:2510.15742 (2025)
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)

3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. arXiv preprint arXiv:2210.09461 (2022)
5. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)
6. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
7. Chen, Y., He, X., Ma, X., Ma, Y.: Contextflow: Training-free video object editing via adaptive context enrichment. arXiv preprint arXiv:2509.17818 (2025)
8. Chen, Z., Zou, Z., Zhang, K., Su, X., Yuan, X., Guo, Y., Zhang, Y.: Dove: Efficient one-step diffusion model for real-world video super-resolution (2025), `https://arxiv.org/abs/2505.16239`
9. Cheng, J., Xiao, T., He, T.: Consistent video-to-video transfer using synthetic dataset. arXiv preprint arXiv:2311.00213 (2023)
10. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems **35**, 16344–16359 (2022)
11. Dong, L., Fan, Q., Guo, Y., Wang, Z., Zhang, Q., Chen, J., Luo, Y., Zou, C.: Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23174–23184 (June 2025)
12. Feng, K., Ma, Y., Wang, B., Qi, C., Chen, H., Chen, Q., Wang, Z.: Dit4edit: Diffusion transformer for image editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2969–2977 (2025)
13. Feng, T., Li, Z., Yang, S., Xi, H., Li, M., Li, X., Zhang, L., Yang, K., Peng, K., Han, S., et al.: Streamdiffusionv2: A streaming system for dynamic and interactive video generation. arXiv preprint arXiv:2511.07399 (2025)
14. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
15. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
16. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
17. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
18. Ju, X., Wang, T., Zhou, Y., Zhang, H., Liu, Q., Zhao, N., Zhang, Z., Li, Y., Cai, Y., Liu, S., et al.: Editverse: Unifying image and video editing and generation with in-context learning. arXiv preprint arXiv:2509.20360 (2025)

19. Kodaira, A., Xu, C., Hazama, T., Yoshimoto, T., Ohno, K., Mitsuhori, S., Sugano, S., Cho, H., Liu, Z., Tomizuka, M., et al.: Streamdiffusion: A pipeline-level solution for real-time interactive generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12371–12380 (2025)
20. Li, R., Haji-Ali, M., Mirzaei, A., Wang, C., Sahni, A., Skorokhodov, I., Siarohin, A., Jakab, T., Han, J., Tulyakov, S., et al.: Egoedit: Dataset, real-time streaming model, and benchmark for egocentric video editing. arXiv preprint arXiv:2512.06065 (2025)
21. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 25690–25699 (2025)
22. Li, W., Pan, W., Luan, P.C., Gao, Y., Alahi, A.: Stable video infinity: Infinite-length video generation with error recycling. arXiv preprint arXiv:2510.09212 (2025)
23. Li, Z., Pun, C.M., Fang, C., Wang, J., Cun, X.: Personalive! expressive portrait image animation for live streaming. arXiv preprint arXiv:2512.11253 (2025)
24. Liang, F., Kodaira, A., Xu, C., Tomizuka, M., Keutzer, K., Marculescu, D.: Looking backward: Streaming video-to-video translation with feature banks. arXiv preprint arXiv:2405.15757 (2024)
25. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)
26. Lin, S., Xia, X., Ren, Y., Yang, C., Xiao, X., Jiang, L.: Diffusion adversarial post-training for one-step video generation. arXiv preprint arXiv:2501.08316 (2025)
27. Liu, J., Wang, X., Lin, Y., Wang, Z., Wang, P., Cai, P., Zhou, Q., Yan, Z., Yan, Z., Shi, Z., et al.: A survey on cache methods in diffusion models: Toward efficient multi-modal generation. arXiv preprint arXiv:2510.19755 (2025)
28. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
29. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In: The Twelfth International Conference on Learning Representations (2023)
30. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15762–15772 (2024)
31. Ma, Y., Cun, X., Liang, S., Xing, J., He, Y., Qi, C., Chen, S., Chen, Q.: Magicstick: Controllable video editing via control handle transformations. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9385–9395. IEEE (2025)
32. Ma, Y., Feng, K., Hu, Z., Wang, X., Wang, Y., Zheng, M., He, X., Zhu, C., Liu, H., He, Y., et al.: Controllable video generation: A survey. arXiv preprint arXiv:2507.16869 (2025)
33. Ma, Y., Feng, K., Zhang, X., Liu, H., Zhang, D.J., Xing, J., Zhang, Y., Yang, A., Wang, Z., Chen, Q.: Follow-your-creation: Empowering 4d creation through video inpainting. arXiv preprint arXiv:2506.04590 (2025)
34. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
35. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. arXiv preprint arXiv:2506.05207 (2025)

36. Ma, Y., Wang, X., Ma, Q., Wang, Q., Zheng, M., Yang, X., Li, H., Zhao, C., Ying, J., Yang, H., et al.: Group editing: Edit multiple images in one go. arXiv preprint arXiv:2603.22883 (2026)
37. Ma, Y., Wang, Z., Ren, T., Zheng, M., Liu, H., Guo, J., Fong, M., Xue, Y., Zhao, Z., Schindler, K., et al.: Fastvmt: Eliminating redundancy in video motion transfer. arXiv preprint arXiv:2602.05551 (2026)
38. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6038–6047 (2023)
39. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15932–15942 (2023)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
42. Schuhmann, C.: Laion-aesthetics (2022), https://laion.ai/blog/laion-aesthetics/
43. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
44. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023), https://arxiv.org/abs/2303.01469
45. Team, D.: Lucy edit: Open-weight text-guided video editing (2025), https://d2drjpuinn46lb.cloudfront.net/Lucy_Edit__High_Fidelity_Text_Guided_Video_Editing.pdf
46. Teng, H., Jia, H., Sun, L., Li, L., Li, M., Tang, M., Han, S., Zhang, T., Zhang, W., Luo, W., et al.: Magi-1: Autoregressive video generation at scale. arXiv preprint arXiv:2505.13211 (2025)
47. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
48. Wang, J., Ma, Y., Guo, J., Xiao, Y., Huang, G., Li, X.: Cove: Unleashing the diffusion feature correspondence for consistent video editing. Advances in Neural Information Processing Systems **37**, 96541–96565 (2024)
49. 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)
50. Wang, J., Lin, S., Lin, Z., Ren, Y., Wei, M., Yue, Z., Zhou, S., Chen, H., Zhao, Y., Yang, C., et al.: Seedvr2: One-step video restoration via diffusion adversarial post-training. arXiv preprint arXiv:2506.05301 (2025)

51. Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., Sang, N.: Videolcm: Video latent consistency model. arXiv preprint arXiv:2312.09109 (2023)
52. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7623–7633 (2023)
53. Yang, S., Xi, H., Zhao, Y., Li, M., Zhang, J., Cai, H., Lin, Y., Li, X., Xu, C., Peng, K., Chen, J., Han, S., Keutzer, K., Stoica, I.: Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation (2025), https://arxiv.org/abs/2505.18875
54. Yang, X., Xie, J., Yang, Y., Huang, Y., Xu, M., Wu, Q.: Unified video editing with temporal reasoner. arXiv preprint arXiv:2512.07469 (2025)
55. 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)
56. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025)
57. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
58. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22963–22974 (2025)
59. Zhang, T., Zhang, Y., Vineet, V., Joshi, N., Wang, X.: Controllable text-to-image generation with gpt-4. arXiv preprint arXiv:2305.18583 (2023)
60. Zheng, S., Feng, L., Wang, X., Zhou, Q., Cai, P., Zou, C., Liu, J., Lin, Y., Chen, J., Ma, Y., et al.: Forecast then calibrate: Feature caching as ode for efficient diffusion transformers. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13449–13457 (2026)
61. Zheng, Z., Wang, X., Zou, C., Wang, S., Zhang, L.: Compute only 16 tokens in one timestep: Accelerating diffusion transformers with cluster-driven feature caching. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 10181–10189 (2025)
62. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)