STREAMINGEFFECT: REAL-TIME HUMAN-CENTRIC VIDEO EFFECT GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: Given an incoming human-centric video stream, StreamingEffect supports interactive real-time control by switching the reference image or text prompt on the fly, enabling stable long-term effects without flickering or drift. Representative frames are shown.

ABSTRACT

Streaming video effect generation is highly desirable for live human-centric applications such as e-commerce streaming, entertainment, and vlogging, yet remains difficult due to the lack of suitable data and deployable editing models. Unlike generic video generation, this task requires real-time video-to-video editing that adds expressive effects while preserving human identity, background content, and temporal consistency. Existing acceleration efforts mainly focus on text-to-video generation, while efficient distillation for video editing remains largely underexplored. In this paper, we present **StreamingEffect**, a real-time human-centric streaming video effect framework. We adopt an in-context video editing architecture and train a high-quality bidirectional teacher, then distill it into a causal autoregressive student and further reduce sampling from 50 steps to 4 steps. We also introduce keyframe control, allowing reference effect frames to be injected online and propagated through the stream for interactive editing. To address the data bottleneck, we construct **VideoEffect-130K**, to our knowledge the largest human-centric video effect dataset, containing 70K effect videos and 60K editing videos across 600 effect categories curated from short-video and editing platforms. Experiments show that our method enables real-time, high-quality 720p video editing on a single H200 GPU.

1 INTRODUCTION

Live video has become a dominant medium for digital interaction, spanning e-commerce streaming, short-video creation, entertainment, and personal vlogging. In these scenarios, visual effects play an important role in improving attractiveness and engagement, by adding accessories, stylized

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

appearances, decorative elements, or atmosphere cues to human-centric videos. Although such effects are widely desired in real applications, current industry systems still largely rely on manually designed assets and detection-tracking-attachment pipelines. These template-based solutions are efficient for predefined effects, but are limited in diversity, semantic flexibility, and visual expressiveness, making it difficult to support open-ended and instruction-driven effect creation.

Streaming video effect generation is therefore a highly demanded but technically challenging problem. Given an incoming video stream, the goal is to continuously generate edited frames with expressive and controllable effects under strict latency constraints. This task differs from conventional offline video editing in three key aspects. First, it requires real-time streaming inference rather than full-sequence generation. Second, it must preserve editing consistency: the human identity, body structure, background, and scene content should remain stable before and after editing, while only the desired effects are added. Third, it requires flexible controllability, allowing effects to be specified not only by text but also by reference images or keyframes for interactive creation.

Despite its practical importance, this problem remains underexplored due to two major gaps. The first is a data gap. Existing video editing datasets (Lin et al., 2026; Wu et al., 2025; Yu et al., 2025) are often small-scale, synthetic, or not tailored to human-centric effect generation. In particular, there is a lack of large-scale paired real-video data that captures how diverse visual effects are applied to faces, bodies, clothing, and surrounding human-centric scenes. To address this gap, we construct **VideoEffect-130K**, a large-scale paired dataset built from both platform-style effect sources and synthetic video editing pipelines, covering real short-video and editing platforms such as TikTok, Douyin, Kuaishou, and Jianying-style effect sources. It contains 130K high-quality paired videos covering 600 effect categories, and is, to our knowledge, the largest dataset specifically designed for human-centric video effect generation.

The second gap lies in efficient video editing deployment. Recent acceleration techniques have made substantial progress in text-to-video generation, including causal generation (Yin et al., 2025; Huang et al., 2025; Lin et al., 2025) and few-step distillation (Salimans & Ho, 2022; Song et al., 2023; Yin et al., 2024b). However, streaming video editing is fundamentally different: the model must preserve the input video content, follow effect conditions, maintain temporal consistency, and generate edited frames with low latency. Efficient distillation for video editing, especially for human-centric effect generation, remains far less explored. Meanwhile, strong image editing models (Brooks et al., 2023; Zhang et al., 2023; Kawar et al., 2023; Hertz et al., 2022) can produce high-quality effect keyframes, but they operate on individual images and cannot directly propagate effects through a video stream.

In this paper, we present **StreamingEffect**, a real-time streaming video-to-video framework for human-centric effect generation. We start from an in-context video editing architecture, where the input video and effect conditions are jointly modeled in a shared context space. This design allows the model to preserve source-video information while adding new effects, which is essential for maintaining identity, background, and temporal consistency. We first train a high-quality bidirectional teacher model on short clips to obtain strong visual quality and temporal coherence. To make the model deployable in streaming scenarios, we adopt a two-stage distillation pipeline inspired by advances in causal video distillation (Yin et al., 2025; Huang et al., 2025): bidirectional-to-causal autoregressive distillation for chunk prediction, followed by multi-step-to-few-step distillation (Song et al., 2023; Yin et al., 2024b; Sauer et al., 2024) that reduces sampling from 50 to 4 steps.

To further improve controllability and visual quality, we introduce keyframe control. Instead of relying only on text prompts, users or upstream image editing models can provide a reference effect keyframe during streaming inference. StreamingEffect propagates the reference effect across frames with temporal coherence, supporting image-driven control and updates. With system-level optimization, it achieves real-time 720p video editing on a single H200 GPU.

Our contributions are summarized in three aspects:

- We introduce **streaming video effect generation**, a practical yet challenging task for live human-centric applications, requiring real-time inference, editing consistency, aesthetic quality, and flexible effect control.
- We propose **StreamingEffect**, a real-time streaming video-to-video framework that integrates in-context video editing, causal autoregressive distillation, few-step acceleration, and keyframe-based interactive control.

- We build **VideoEffect-130K**, a 130K-pair dataset covering 600 effect and editing categories, combining rendered effect data with synthetic/general-editing data for controllable human-centric video effect generation.

2 Related Work

Video Generation Models. Recent video generation has been driven by diffusion- and transformer-based foundation models built on image diffusion priors (Ho et al., 2020; Song et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023). Early landmark video diffusion systems including *VDM* (Ho et al., 2022b), *Imagen Video* (Ho et al., 2022a), *Make-A-Video* (Singer et al., 2022), *Video LDM* (Blattmann et al., 2023b), *AnimateDiff* (Guo et al., 2023), and *Lumiere* (Bar-Tal et al., 2024) establish the recipe of jointly modeling spatial and temporal context, while commercial-scale systems such as *Sora* (Brooks et al., 2024) demonstrate emerging large-scale capability. *Stable Video Diffusion* further shows that scaling latent video diffusion with curated data produces strong text-to-video and image-to-video synthesis priors (Blattmann et al., 2023a), and recent open systems including *Latte* (Ma et al., 2024a), *Emu Video* (Girdhar et al., 2023), *VideoCrafter2* (Chen et al., 2024b), *CogVideoX* (Yang et al., 2024), *HunyuanVideo* (Kong et al., 2024), and *Wan* (Wan et al., 2025) continue to push motion quality, semantic alignment, duration, and fidelity through stronger spatiotemporal representations, text-video fusion, and large-scale training recipes. Beyond unconditional or text-driven synthesis, controllable video generation further incorporates visual, structural, motion, trajectory, layout, or reference-based conditions to improve user control over generated content (Jiang et al., 2025; Zhang et al., 2025; Song et al., 2025a; 2024; Ma et al., 2024b; 2025a; 2024c; 2025b; Song et al., 2025b; Yang et al., 2025; Song et al., 2026). These models have advanced open-source video generation, but they are designed for offline creation where quality is prioritized over low-latency streaming. Thus, they provide backbones but do not address real-time, human-centric video effect editing.

Real-Time Video Generation. Video diffusion models remain costly due to large model sizes, full-sequence attention, and multi-step denoising. Building on a long line of step-distillation methods originally developed for image diffusion—progressive distillation (Salimans & Ho, 2022), consistency models (Song et al., 2023; Luo et al., 2023), distribution-matching distillation (Yin et al., 2024b;a), adversarial diffusion distillation (Sauer et al., 2024), rectified flow (Liu et al., 2022), and one-step generators (Liu et al., 2024b)—recent work pushes video efficiency through compressed latent modeling, streaming architectures, and distillation. *LTX-Video* operates in highly compressed latent spaces, while *StreamDiT* introduces streaming generation with moving buffers (Kodaira et al., 2025). Distillation-based methods including *DOLLAR* (Ding et al., 2025), *Diagonal Distillation* (Liu et al., 2026), *Rolling Forcing* (Liu et al., 2025a), *CausVid* (Yin et al., 2025), *Self-Forcing* (Huang et al., 2025), and *APT* (Lin et al., 2025) reduce sampling steps or convert bidirectional/diffusion generators into causal or long-horizon autoregressive models. These works motivate combining autoregressive streaming with step reduction, but mainly target generic video generation rather than human-centric video editing, where source preservation, effect consistency, and interactivity remain critical.

Video Editing and Visual Effects. Video editing has progressed from tuning-free adaptations of image diffusion models (Wu et al., 2023; Qi et al., 2023; Ceylan et al., 2023; Liu et al., 2024a; Cheng et al., 2023; Song et al., 2025c; Guo et al., 2025; Zhang et al., 2024; Wang et al., 2025; Gong et al., 2025) to video-native editing frameworks. *Rerender A Video*, *TokenFlow*, and *CoDeF* exploit image diffusion priors with temporal correspondence, feature propagation, or canonical-field representations for coherent text-guided editing (Yang et al., 2023; Geyer et al., 2023; Ouyang et al., 2024; Song et al., 2024). *AnyV2V* decomposes video editing into first-frame image editing—enabled by strong image editors such as *InstructPix2Pix* (Brooks et al., 2023), *ControlNet* (Zhang et al., 2023), *Imagic* (Kawar et al., 2023), and *Prompt-to-Prompt* (Hertz et al., 2022)—followed by image-to-video generation, while *VACE* unifies video creation and editing through a video-native framework (Ku et al., 2024; Jiang et al., 2025). *StreamV2V* studies online video-to-video translation with feature banks, but focuses on general prompt-based translation rather than effect-centric human video editing (Liang et al., 2024) (Lin et al., 2026) (Pan et al., 2026). For visual effects, *VFX Creator* (Liu et al., 2025b), *VFXMaster* (Li et al., 2025a), *Transanimate* (Chen et al., 2025), and *IC-Effect* (Li et al., 2025b) explore controllable VFX generation with mask control, reference videos, or in-context conditioning. In contrast, our work targets real-time human-centric streaming video effect generation, emphasizing low-latency deployment, source-video preservation, and online keyframe control.

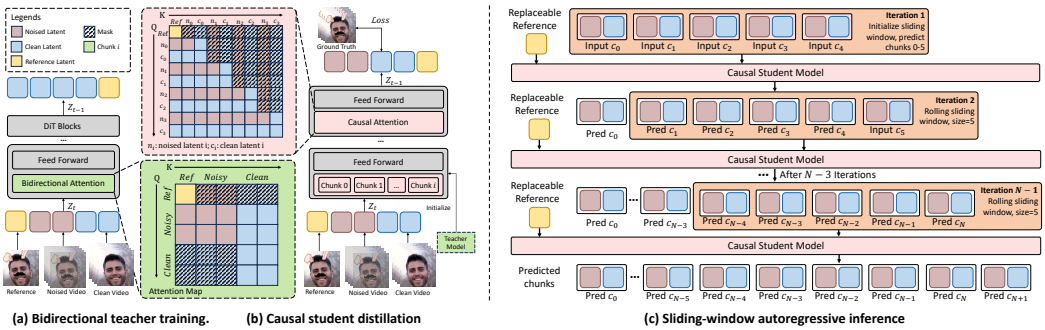


Figure 2: **Overview of StreamingEffect.** (a) **Bidirectional Teacher Training:** A bidirectional teacher is trained with reference-conditioned in-context video editing. (b) **Causal Student Distillation:** The teacher is distilled into a causal autoregressive student for streaming generation. (c) **Sliding-window Autoregressive Inference:** The student edits incoming video chunks online with cached context and propagates effects through the stream.

3 METHOD

We present **StreamingEffect**, a real-time streaming video-to-video framework for human-centric effect generation. Sec. 3.1 defines the streaming formulation, Sec. 3.2 introduces our reference-conditioned in-context editing architecture, Sec. 3.3 trains a high-quality bidirectional teacher, and Sec. 3.4 converts it into a causal few-step streaming student. Finally, Sec. 3.5 describes the construction of VideoEffect-130K.

3.1 PROBLEM DEFINITION

We study *streaming video effect generation*, a human-centric video-to-video editing task that adds expressive effects to an incoming video stream under strict real-time constraints. The effects considered in this work include global or local stylization, accessories, headwear, makeup, decorative objects, and other effects commonly used in live commerce, entertainment, and creator scenarios. The task requires not only visually appealing effect rendering, but also strong preservation of human identity, body structure, background content, and temporal consistency.

Let $\mathbf{X} = \{x_t\}_{t=1}^{T}$ denote a source video and $\mathbf{Y} = \{y_t\}_{t=1}^{T}$ denote the target video after applying the desired effect. We introduce a reference image $\mathbf{r}$ as an image-level effect condition, with a text prompt $\mathbf{c}_{\text{txt}}$ defining the desired effect. Our goal is to learn a streaming generator G that maps the source video and reference condition to the edited video:

$$\hat{\mathbf{Y}} = G(\mathbf{X}, \mathbf{r}, \mathbf{c}_{\text{txt}}). \tag{1}$$

For streaming inference, we divide the video into temporal chunks, $\mathbf{X} = \{\mathbf{X}_0, \ldots, \mathbf{X}_M\}$ and $\mathbf{Y} = \{\mathbf{Y}_0, \ldots, \mathbf{Y}_M\}$, and generate each target chunk causally:

$$\hat{\mathbf{Y}}_i = G(\mathbf{X}_{\le i}, \hat{\mathbf{Y}}_{<i}, \mathbf{r}, \mathbf{c}_{\text{txt}}), \quad i = 0, \ldots, M. \tag{2}$$

At deployment, past predictions enter G through cached keys and values (Sec. 3.4), and a sliding window bounds per-chunk cost over long streams. Together with chunk size and few-step denoising, this gives a constant per-chunk latency budget.

3.2 REFERENCE-CONDITIONED IN-CONTEXT VIDEO EDITING

StreamingEffect is built on a pretrained Wan2.2-TI2V-5B backbone (Wan et al., 2025) and adopts a reference-conditioned in-context video editing architecture. Instead of injecting the source video through a separate control branch, we encode the source video, target video, and reference image into a shared latent token space and organize them into one context sequence. This design allows the denoising transformer to directly model the relationship among the original video content, the reference effect, and the target edited frames, which is important for preserving identity, background, and scene structure while adding new effects.

Let $\mathbf{z}^x$, $\mathbf{z}^y$, and $\mathbf{z}^r$ denote the latent tokens of the source video, target video, and reference image. We arrange them into an interleaved sequence:

$$[\mathbf{z}^r \mid \mathbf{z}_1^x \mid \mathbf{z}_1^y \mid \mathbf{z}_2^x \mid \mathbf{z}_2^y \mid \cdots \mid \mathbf{z}_M^x \mid \mathbf{z}_M^y]. \tag{3}$$

Each pair $(\mathbf{z}_i^x, \mathbf{z}_i^y)$ forms a temporal super-chunk $\mathcal{S}_i$. A bidirectional model uses full attention over the sequence for high-quality short-clip editing, while the final streaming student replaces full attention with block-causal attention and KV caching for online generation.

The reference image $\mathbf{r}$ provides keyframe-level effect control. During training we sample $\mathbf{r}$ uniformly at random from the frames of $\mathbf{Y}$, forcing the model to learn how a single edited keyframe should guide effect rendering and temporal propagation. During inference, $\mathbf{r}$ can be obtained by applying a strong image editing model or commercial image-editing API to a selected source frame. StreamingEffect then propagates this high-quality reference effect to subsequent frames in a temporally coherent manner. This design allows our streaming video editor to benefit from the strong visual quality of modern editing models while maintaining video-level consistency and low latency.

3.3 Bidirectional Teacher Training

We first train a bidirectional teacher model F_ϕ using the in-context formulation above. The teacher receives the source video, the target video, and the reference image condition, and is optimized for visual quality and temporal coherence without enforcing streaming causality. Given a training tuple $(\mathbf{X}, \mathbf{Y}, \mathbf{r})$, the source video, target effect video, and sampled reference frame are encoded into latent tokens and processed by a transformer denoiser with full bidirectional attention. This allows each target token to attend to both past and future context, which is beneficial for effects that depend on face motion, body pose evolution, delayed appearance changes, or long-range temporal consistency.

We train the teacher with the native rectified-flow objective (Esser et al., 2024) of the backbone. Let $\mathbf{z}^y$ be the clean target latent and $\mathbf{z}_t^y$ its noisy interpolation at time t. The teacher F_ϕ predicts the flow direction conditioned on the source and reference latents:

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}\left[\|F_\phi(\mathbf{z}_t^y, \mathbf{z}^x, \mathbf{z}^r, \mathbf{c}_{\text{txt}}, t) - \mathbf{v}^\star\|_2^2\right], \tag{4}$$

where $\mathbf{v}^\star$ denotes the target flow field. We apply condition dropout during training on the semantic conditions $(\mathbf{z}^r, \mathbf{c}_{\text{txt}})$, encouraging the teacher to learn each as an independent signal. The reference image is dropped at rate 0.5 and the text at rate 0.1. When dropped, $\mathbf{z}^r$ is replaced with an all-zero latent of identical shape. This further enables classifier-free guidance (Ho & Salimans, 2022) at inference. The trained teacher provides strong effect fidelity and temporal coherence, but relies on full-sequence attention and many denoising steps, making it unsuitable for real-time streaming.

3.4 Two-Stage Distillation for Streaming Inference

Converting the bidirectional teacher into a real-time streaming editor requires more than swapping full attention with causal attention. The student must handle heterogeneous noise levels across time, as past chunks become clean and cached while the current chunk is still being denoised. Moreover, it must be trained on its own rollout distribution to reduce error accumulation over long streams. We therefore use a two-stage distillation pipeline: Stage 1 performs causal adaptation with diffusion forcing (Chen et al., 2024a), and Stage 2 performs on-policy self-forcing (Huang et al., 2025) for few-step streaming generation to close the exposure-bias gap.

Stage 1: Bidirectional-to-Causal Adaptation. The first stage converts the bidirectional teacher into a causal autoregressive student. We keep the same interleaved super-chunk layout of Eq. (3) but replace full attention with a block-causal mask: each super-chunk $\mathcal{S}_i$ attends bidirectionally within itself and to the reference $\mathbf{z}^r$ and all previous super-chunks $\mathcal{S}_{j\le i}$, but not to future chunks. At inference time, a persistent KV cache stores keys and values from previous chunks, avoiding repeated computation and enabling low-latency streaming generation.

Stage 1 also adapts the timestep distribution. Standard diffusion training samples a single timestep for the whole video, while streaming inference contains mixed noise states: previous chunks are already clean in the cache, whereas the current chunk remains noisy. We therefore assign each super-chunk an independent timestep t_i, including the clean state $t_i{=}0$, so the student observes heterogeneous

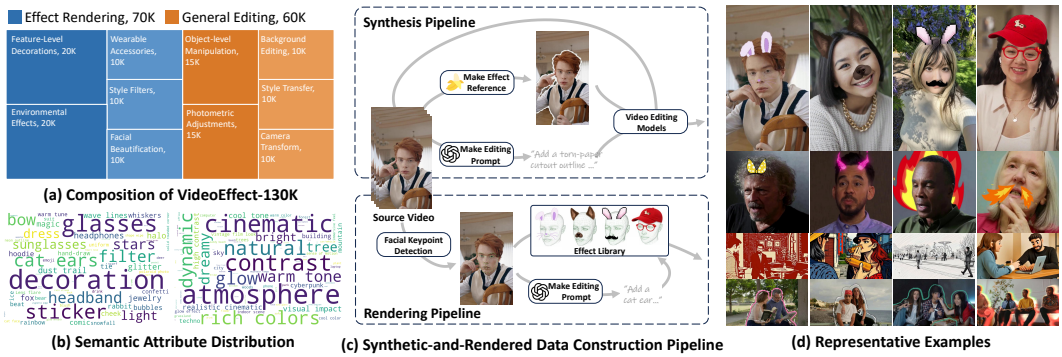


Figure 3: **Construction and statistics of VideoEffect-130K.** The dataset contains 130K paired human-centric videos, including 70K rendered effect samples and 60K general editing samples across about 600 categories. It is built with a hybrid synthetic-and-rendered pipeline, and each sample consists of a source video, a reference effect image, and a target edited video.

noise configurations encountered during autoregressive generation. The student is initialized from the teacher ($G_\theta \leftarrow F_\phi$) and optimized with the same rectified-flow denoising objective under the causal mask. This stage enables chunk-wise streaming generation, but still relies on teacher-forced ground-truth latents and thus retains rollout-induced exposure bias.

Stage 2: On-Policy Self-Forcing for Few-Step Generation. The second stage closes the few-step gap and the exposure-bias gap simultaneously. Following Self Forcing (Huang et al., 2025), we roll out the Stage-1 student exactly as it will run at inference: across chunks autoregressively, with each new chunk attending to a KV cache $\mathrm{KV}_{<i}$ populated by the student's own previous-chunk predictions $\hat{\mathbf{z}}^y_{<i}$. Within each chunk, we simulate the deployment-time K-step rectified-flow sampler under classifier-free guidance: at each rollout step k for chunk i, the student performs two forward passes that share the same cache. A conditional one $\mathbf{v}_{\text{cond}} = G_\theta(\mathbf{z}^y_{i,t_k}, \mathbf{z}^x, \mathbf{z}^r, \mathbf{c}_{\text{txt}}; \mathrm{KV}_{<i}, t_k)$ and one with both semantic conditions ablated $\mathbf{v}_{\text{uncond}} = G_\theta(\mathbf{z}^y_{i,t_k}, \mathbf{z}^x, \varnothing, \varnothing; \mathrm{KV}_{<i}, t_k)$. The two are blended at the teacher's CFG scale w to produce $\tilde{\mathbf{v}}_{t_k} = (1{+}w)\mathbf{v}_{\text{cond}} - w\mathbf{v}_{\text{uncond}}$ and the $\mathbf{x}_0$-prediction $\hat{\mathbf{z}}^y_{i,0,k}$ that drives the rollout to step $k{+}1$. Gradients flow only through the conditional forward and the unconditional forward is detached.

Because our task is conditional video-to-video editing with paired data, each $(\mathbf{X}, \mathbf{r}, \mathbf{c}_{\text{txt}})$ corresponds to a specific target $\mathbf{Y}$. This allows us to supervise the on-policy rollout directly with ground-truth target latents $\mathbf{z}^y$, rather than relying only on teacher scores, reward models, or discriminators (Yin et al., 2024b;a). We use an SNR-weighted regression objective over the rollout:

$$\mathcal{L}_{\text{stage2}} = \frac{1}{\sum_k w_k} \sum_{k=1}^{K} w_k \left\| \hat{\mathbf{z}}^y_{0,k} - \mathbf{z}^y \right\|_2^2, \quad w_k = \frac{(1-t_k)^2}{t_k^2}. \tag{5}$$

SNR weighting emphasizes low-noise steps where the clean target is more strongly determined. This is key to stabilizing ultra-low-step generation and reducing error accumulation across chunks.

After Stage 2, the final student performs causal streaming inference with KV-cached attention and only a few denoising steps. In our implementation, the sampling process is reduced from 50 steps to 4 steps, enabling real-time 720p video effect generation while retaining much of the bidirectional teacher's visual quality and temporal coherence.

3.5 Dataset Construction

To support streaming video effect generation, we construct **VideoEffect-130K**, a large-scale paired video dataset for human-centric effect rendering and general video editing. As shown in Figure 3, the dataset contains about 130K source-target video pairs, including **70K effect-rendering samples** and **60K general-editing samples**. The effect-rendering subset covers feature-level decorations, environmental effects, wearable accessories, style filters, and facial beautification, while the general-editing subset includes object-level manipulation, photometric adjustments, background editing, style transfer, and camera transformation.

VideoEffect-130K is built through a hybrid synthetic-and-rendered pipeline. The *synthesis* branch uses source videos, generated prompts, effect references, and video editing models to produce diverse target videos, improving coverage of open-ended editing instructions. The *rendering* branch applies a platform-style effect library to source videos, optionally with facial keypoints and other human-centric cues for spatial alignment, providing stable supervision for common live-streaming effects such as facial accessories, headwear, makeup-like effects, local decorations, and atmosphere effects.

After construction, we apply filtering, trimming, categorization, and quality control to remove low-quality, temporally unstable, or poorly aligned pairs, and preprocess raw videos into 720p clips. Each sample is organized as a triplet of source video, reference image, and target video. During training, the reference image is randomly sampled from the target video, encouraging the model to learn reference-keyframe-conditioned effect rendering and temporal propagation over the video stream.

4 EXPERIMENTS

4.1 SETUP

Implementation Details. StreamingEffect is built on the pretrained *Wan2.2-TI2V-5B* backbone. All training and evaluation are performed at 720p with multi-aspect-ratio bucketed inputs (1:1, 9:16, 4:3, etc.), each clip containing 97 pixel frames at 24 fps encoded into 25 latent frames and grouped into 5 super-chunks (chunk size 5). The bidirectional teacher is trained on 8 H200 GPUs with LoRA, batch size 4 per GPU, learning rate 1×10^{-4}, weight decay 1×10^{-2}, *AdamW*, for 8,000 steps. **Stage 1 (Bidirectional→Causal SFT)** initializes the causal AR student from the teacher and trains with FSDP, batch size 4 per GPU, learning rate 5×10^{-5}, weight decay 1×10^{-2}, CFG scale 5.0, for 3,000 iterations; per-chunk independent timesteps are sampled uniformly in $[0.001, 0.999]$ to expose the student to heterogeneous noise configurations. **Stage 2 (On-Policy Self-Forcing, 4-step)** starts from the Stage-1 checkpoint and runs on-policy rollouts with batch size 4 per GPU, student learning rate 1×10^{-6}, weight decay 1×10^{-2}, for 3,000 iterations, using a 4-step shifted schedule with timestep list $[0.999, 0.937, 0.833, 0.624, 0.0]$.

Baselines. We compare against two commercial systems, *Kling O1* and *Kling 3.0 Omni* (Kling Team, 2025), accessed via public APIs, and two open-source video-editing models, *Kiwi-Edit* (Lin et al., 2026) and *OmniWeaving* (Pan et al., 2026), all evaluated under 720p, 97-frame configuration. None of these baselines support online streaming inference, so we compare against their offline outputs to assess effect quality independently of latency.

Benchmarks. We evaluate on a held-out test split of 2,000 samples drawn from the 70K effect-rendering subset of VideoEffect-130K, covering around 600 diverse visual effect categories. Each test sample consists of a triplet of (source video, reference effect image, target edited video), with 97-frame 720p clips at multiple aspect ratios.

Evaluation Metrics. We report full-reference metrics MSE, PSNR, SSIM, and LPIPS computed frame-by-frame between the prediction and the ground-truth target. FPS is measured on a single NVIDIA H200 GPU at 720×1280 resolution with each method's default sampling configuration, and is unavailable for the API-only commercial baselines; since FPS is determined by model architecture and sampling schedule rather than the conditioning modality, it does not change between image-conditioned and text-driven settings. Since effect generation is open-ended and many plausible edits satisfy the same instruction, we additionally use a VLM-based protocol in which *Gemini 2.5 Pro* (Comanici et al., 2025) and *GPT-4o* (OpenAI, 2023) independently score each video on a 1–10 scale along four dimensions averaged across the two models: **V-Cons.** (temporal consistency), **E-Ref.** (effect-reference alignment), **Text** (text-following), and **Quality** (overall editing quality, including identity and background preservation).

4.2 QUALITATIVE AND QUANTITATIVE EVALUATION

Figures 4 and 5 show our qualitative results. Compared to StreamingEffect, the open-source baselines tend to lose human identity and exhibit flicker on faces and clothing, especially under motion. Commercial systems achieve stabler structure, but still suffer from identity drift and occasional loss of detailed reference effects. In contrast, our model consistently preserves identity, pose, and background while propagating the reference effect across the full clip with better temporal stability.



Figure 4: **Qualitative results of StreamingEffect.** (a) **Image-guided:** propagating reference-image effects while preserving identity. (b) **Text-guided:** following text prompts to generate diverse visual effects. Representative frames are shown.

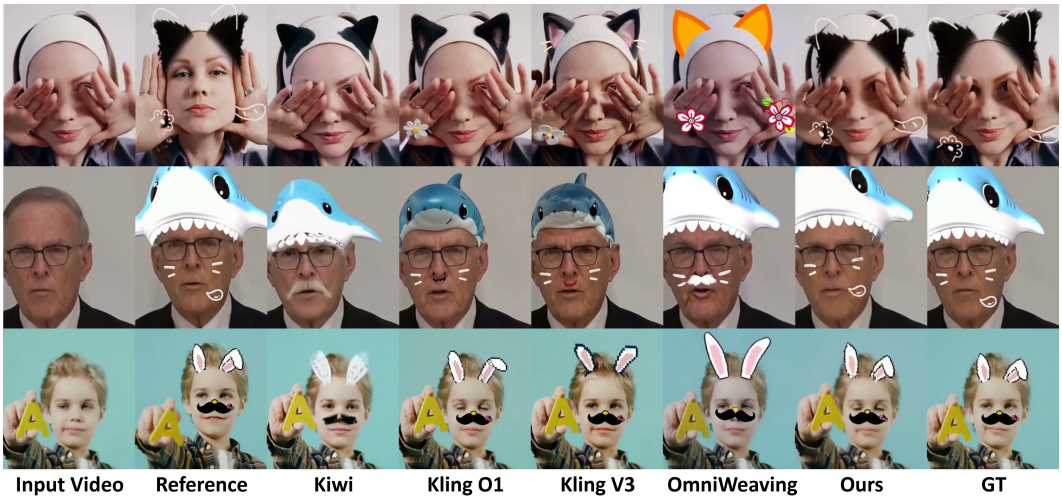


Figure 5: Qualitative comparison against open-source (Kiwi-Edit, OmniWeaving) and commercial (Kling O1, V3) baselines. Representative frames are shown.

Table 1 reports the main comparison under the image-conditioned setting. Our bidirectional teacher leads on full-reference fidelity, achieving better MSE, PSNR, and LPIPS than the baselines, and also obtains the best scores on all four VLM dimensions. These results suggest that the teacher can better preserve the target video structure while faithfully transferring the reference effect. The gains are most evident on effect-reference alignment and overall quality, indicating stronger reference-conditioned effect rendering and more stable human-centric editing. Open-source baselines lag behind the commercial systems on E-Ref. and Quality, showing that faithful reference-effect transfer to human-centric video remains an open challenge.

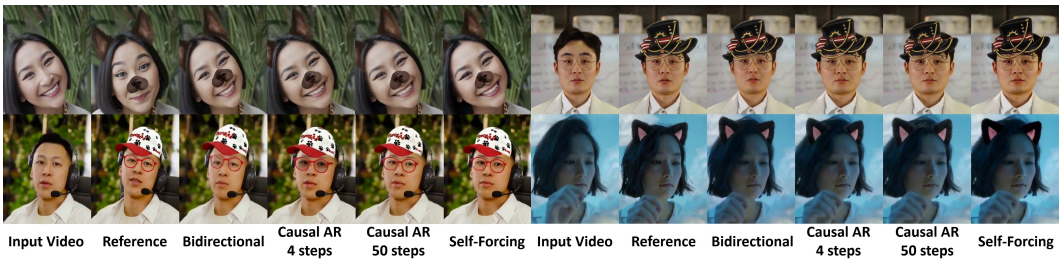


Figure 6: **Qualitative ablation across distillation stages. Bidirectional**: 50-step teacher with full attention. **Causal AR 4/50 steps**: Stage-1 causal student at 4 or 50 denoising steps. **Self-Forcing**: our final Stage-2 student (4 steps). Representative frames are shown.

Table 1: Quantitative comparison under **image-conditioned** and **text-driven** settings. V-Cons., E-Ref., Text, and Quality are VLM scores for temporal consistency, effect-reference alignment, text-following, and overall quality. Lower is better for MSE/LPIPS; higher otherwise. Best in **bold**.

Method	Image-Guided									Text-Guided		
	MSE↓	PSNR↑	SSIM↑	LPIPS↓	FPS↑	V-Cons.↑	E-Ref.↑	Text↑	Quality↑	V-Cons.↑	Text↑	Quality↑
Kiwi-Edit	0.0172	18.7	0.727	0.227	**1.46**	7.90	6.73	6.94	6.96	7.72	7.12	6.81
Kling O1	0.0173	18.9	**0.782**	0.216	–	8.10	8.18	8.35	7.82	8.05	8.43	7.74
Kling 3.0 Omni	0.0184	18.5	0.723	0.240	–	8.78	8.43	8.73	8.24	8.61	8.62	8.15
OmniWeaving	0.0284	16.7	0.732	0.268	0.05	7.82	6.79	7.41	7.09	7.65	7.58	6.94
Ours (Teacher)	**0.0154**	**19.4**	0.763	**0.195**	0.14	**8.98**	**9.02**	**8.75**	**8.41**	**8.84**	**8.68**	**8.19**

Table 2: Ablation of the streaming distillation pipeline at 720×1280 on a single NVIDIA H200. Lower is better for MSE/LPIPS; higher otherwise. Best in **bold**.

Method	Steps	Image-Conditioned									Text-Driven		
		MSE↓	PSNR↑	SSIM↑	LPIPS↓	FPS↑	V-Cons.↑	E-Ref.↑	Text↑	Quality↑	V-Cons.↑	Text↑	Quality↑
Bidirectional Teacher	50	**0.0154**	**19.4**	**0.763**	**0.195**	0.14	**8.98**	**9.02**	**8.75**	**8.41**	**8.84**	**8.68**	**8.19**
Causal AR Student	50	0.0261	16.6	0.678	0.276	0.63	8.10	7.82	8.16	7.71	7.98	8.05	7.52
Causal AR Student	4	0.0263	16.6	0.659	0.278	7.52	7.93	7.47	7.63	7.26	7.81	7.49	7.08
StreamingEffect	4	0.0251	16.7	0.666	0.269	**14.1**	8.37	8.14	8.55	7.75	8.26	8.38	7.58

4.3 Ablation Study

Effect of streaming distillation. We ablate the two distillation stages in Table 2. The bidirectional teacher offers the strongest quality but its full attention and 50-step sampling give well below real-time throughput at 720×1280. The Stage-1 causal AR student enables streaming via KV caching, yet naively reducing it to 4 steps causes clear degradation across all VLM dimensions, confirming that ultra-low-step sampling is not free for a model trained only with multi-step targets. Our final Stage-2 self-forcing student recovers most of this quality at 4 steps while delivering an order-of-magnitude FPS gain over the teacher. This confirms that causalization enables online generation, while on-policy self-forcing is essential for making ultra-low-step streaming inference practical.

5 Conclusion

We presented **StreamingEffect**, an efficient real-time framework for human-centric streaming video effect generation. The task aims to add expressive and controllable effects to live video streams while preserving identity, background content, and temporal consistency under strict latency constraints. Our framework combines reference-conditioned in-context video editing with a two-stage distillation pipeline, converting a high-quality bidirectional teacher into a causal few-step streaming student. We further introduced keyframe control, which allows strong image editing models to provide high-quality reference effects that are propagated online. To support this task, we constructed **VideoEffect-130K**, a large-scale paired dataset that combines rendered effect data with synthetic/general-editing data for effect-centric human video editing. Experiments show that StreamingEffect enables real-time, high-quality 720p video editing with strong visual fidelity on a single H200 GPU, suggesting a practical direction for deployable generative video effects.

AI USE STATEMENT

Generative video editing models are used in the synthetic-data branch of VideoEffect-130K, as described in Section 3.5. Gemini 2.5 Pro and GPT-4o are used as automated perceptual judges, with inputs, scoring criteria, and aggregation described in Appendix B. Claude Code assisted with portions of the experimental workflow and automated evaluation. The authors reviewed the resulting outputs and key code components and take full responsibility for the code, experimental results, and conclusions. Codex was used to adapt the LaTeX source to the ICLR template and check compilation and layout.

REPRODUCIBILITY STATEMENT

Section 3 describes the architecture, training objectives, streaming distillation, and dataset construction. The experimental setup specifies the backbone, hardware, optimization settings, and evaluation protocol. Appendix B provides the inputs and scoring criteria for the LLM-based evaluation. The complete implementation has been released as open-source project. To preserve double-blind anonymity, we omit the public repository link from this submission.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23206–23217, 2023.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024a. URL https://arxiv.org/abs/2407.01392.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7310–7320, 2024b.

Xuewei Chen, Zhimin Chen, and Yiren Song. Transanimate: Taming layer diffusion to generate rgba video. *arXiv preprint arXiv:2503.17934*, 2025.

Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*, 2023.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17961–17971, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Chuang Zhang, and Jiaming Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

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

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022b.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6007–6017, 2023.

Kling Team. Kling-Omni Technical Report, 2025. URL https://arxiv.org/abs/2512.16776.

Akio Kodaira, Tingbo Hou, Ji Hou, Markos Georgopoulos, Felix Juefei-Xu, Masayoshi Tomizuka, and Yue Zhao. Streamdit: Real-time streaming text-to-video generation. *arXiv preprint arXiv:2507.03745*, 2025.

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

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

Baolu Li, Yiming Zhang, Qinghe Wang, Liqian Ma, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Zhenfei Yin, Yunzhi Zhuge, Huchuan Lu, et al. Vfxmaster: Unlocking dynamic visual effect generation via in-context learning. *arXiv preprint arXiv:2510.25772*, 2025a.

Yuanhang Li, Yiren Song, Junzhe Bai, Xinran Liang, Hu Yang, Libiao Jin, and Qi Mao. Ic-effect: Precise and efficient video effects editing via in-context learning. *arXiv preprint arXiv:2512.15635*, 2025b.

Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. *arXiv preprint arXiv:2405.15757*, 2024.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

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

Jinxiu Liu, Xuanming Liu, Kangfu Mei, Yandong Wen, Ming-Hsuan Yang, and Weiyang Liu. Streaming autoregressive video generation via diagonal distillation. *arXiv preprint arXiv:2603.09488*, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025a.

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

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

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024b.

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

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024a.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4117–4125, 2024b.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024c.

Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025a.

Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6018–6026, 2025b.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8089–8099, 2024.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15932–15942, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

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

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

Quanjian Song, Yiren Song, Kelly Peng, Yuan Gao, and Mike Zheng Shou. Worldwander: Bridging egocentric and exocentric worlds in video generation. *arXiv preprint arXiv:2511.22098*, 2025a.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Yiren Song, Shijie Huang, Chen Yao, Xiaojun Ye, Hai Ci, Jiaming Liu, Yuxuan Zhang, and Mike Zheng Shou. Processpainter: Learn painting process from sequence data. *arXiv preprint arXiv:2406.06062*, 2024.

Yiren Song, Cheng Liu, Weijia Mao, and Mike Zheng Shou. Mitty: Diffusion-based human-to-robot video generation. *arXiv preprint arXiv:2512.17253*, 2025b.

Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025c.

Yiren Song, Xiyao Deng, Pei Yang, Yihan Wang, and Mike Zheng Shou. Omnihumanoid: Streaming cross-embodiment video generation with paired-free adaptation. *arXiv preprint arXiv:2605.12038*, 2026.

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

Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633, 2023.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. InsViE-1M: Effective instruction-based video editing with elaborate dataset construction. *arXiv preprint arXiv:2503.20287*, 2025.

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. X-humanoid: Robotize human videos to generate humanoid videos at scale. *arXiv preprint arXiv:2512.04537*, 2025.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. VEGGIE: instructional editing and reasoning video concepts with grounded generation. *arXiv:2503.14350*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8069–8078, 2024.

Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19513–19524, 2025.

A LIMITATIONS

A primary limitation of our work stems from the composition of VideoEffect-130K. Both the platform-rendered subset and the synthetic-editing subset are sourced from short-video and live-streaming platforms whose creator content is overwhelmingly human-centric, so the dataset is dominated by face, body, and portrait scenes. As a consequence, StreamingEffect generalizes well to human-centric effects—accessories, makeup, headwear, portrait stylization, and atmosphere overlays—but is noticeably weaker on non-human scenarios such as pure object close-ups, animals, landscapes, and product shots, where it can miss the intended effect or apply it to the wrong region. Extending the dataset with non-human effect categories, and training a more general streaming editor, is a natural direction for future work.

B LLM-BASED EVALUATION DETAILS

We employ two strong multimodal large language models, Gemini 2.5 Pro (Comanici et al., 2025) and GPT-4o (OpenAI, 2023), as automated perceptual judges for human-centric video effect generation. To reduce single-judge bias and scoring variance, each metric is queried three times per model to obtain a median score, and the final reported score is the average of the medians from both Gemini 2.5 Pro and GPT-4o.

Inputs. For every test sample, the judges receive (i) a horizontally concatenated video with three columns—left: the input/source video, middle: the generated video, right: the ground-truth target video—(ii) a reference effect image, and (iii) a text prompt describing the desired effect. The judges are explicitly instructed to score only the generated video in the middle column, while using the source video, the ground-truth target video, the reference image, and the text prompt as evaluation context. They are asked to return JSON of the form `{"score": <1-10>, "reason": "<sentence>"}` for each of four metrics: **Video Temporal Consistency** (`V-Cons.`), **Effect–Reference Consistency** (`E-Ref.`), **Overall Editing Quality** (`Quality`), and **Text-Following Ability** (`Text`). Aggregated results are reported in Table 1 and Table 2 of the main paper.

Video Temporal Consistency (`V-Cons.`). The judges evaluate whether the generated video is temporally stable across frames. Criteria include the absence of flickering and jitter, no sudden appearance or disappearance of effects, stable identity throughout the clip, consistent background, natural and continuous motion, and the absence of temporal artifacts. Score 10 indicates a highly stable and coherent result over time; 8–9 minor temporal artifacts; 6–7 noticeable flicker or jitter; 4–5 strong instability; 1–3 severe flickering or temporal breakdown.

Effect–Reference Consistency (`E-Ref.`). The judges evaluate whether the generated effect matches the provided reference effect image. Criteria include the consistency of appearance, style, color, accessory, makeup, decoration, or visual effect with the reference image, and whether the effect is propagated correctly and stably over time. Score 10 indicates the generated effect is highly consistent with the reference; 8–9 minor deviation in style or color; 6–7 the effect is recognizable but partially mismatched; 4–5 only loosely related to the reference; 1–3 the effect is missing, unrelated, or clearly inconsistent with the reference image.

Overall Editing Quality (`Quality`). The judges evaluate the overall visual quality of the edited video. Criteria include realism and sharpness, identity preservation of the subject, preservation of body structure, preservation of the background, natural blending of the effect with the underlying scene, the absence of artifacts, and overall aesthetic quality. Pixel-level identity to the ground truth is explicitly not required, as long as the intended edit is achieved with high perceptual quality. Score 10 indicates excellent editing quality with natural and clean results; 8–9 high quality with minor artifacts; 6–7 acceptable but with visible flaws; 4–5 noticeable distortion or blending issues; 1–3 poor visual quality, severe artifacts, identity distortion, or failed editing.

Text-Following Ability (`Text`). The judges evaluate whether the generated video follows the provided text prompt. Criteria include whether the requested effect type, object, style, location, and semantic instruction are correctly reflected in the generated video. Score 10 indicates the generated video fully follows the text prompt; 8–9 the prompt is largely followed with minor omissions; 6–7

the main intent is captured but secondary instructions are missed; 4–5 the prompt is only partially followed; 1–3 the generated video ignores or contradicts the text prompt.

Scoring rules. The judges are instructed to be strict but fair, and to follow several explicit rules: focus only on the generated video in the middle column; use the source video to judge content preservation; treat the ground-truth target video as a reference for the intended edit rather than a pixel-level target; use the reference image to judge effect consistency; and use the text prompt to judge semantic instruction following. High scores are not given when the effect looks visually pleasing but the subject's identity, body structure, or background are altered incorrectly; conversely, low scores are not given solely for slight deviations from the ground truth when the intended effect is correct and visually plausible.