

OSVE: One Step Video Editing with One Step Diffusion Models

Habin Lim[✉] and Gyeong-Moon Park^{✉*}

Korea University, Seoul, Republic of Korea
{ha001211, gm-park}@korea.ac.kr

Abstract. Text-guided video editing with diffusion models is impractically slow, hindered by costly multi-step sampling and inversion. We present OSVE, the first framework to successfully adapt one-step Text-to-Image (T2I) models for high-quality video editing, addressing the core challenges of inversion, editability, and temporal consistency. To bypass slow iterative inversion, we train a learnable encoder that predicts the initial noise for each frame in a single forward pass. This encoder is trained with a novel Structure-Aware Editing (SAE) loss on a curated dataset of structurally-aligned image pairs, teaching it to preserve the source video’s geometry during edits. For temporal coherence, we introduce Unified-Frame Editing (UFE), a technique that concatenates frame latents to facilitate cross-frame attention in a single generation step. Furthermore, for long videos, a sliding-window strategy with an anchor frame maintains global consistency. Our extensive experiments demonstrate that OSVE achieves editing quality comparable or superior to state-of-the-art multi-step methods, while operating approximately 155–171 times faster. This breakthrough paves the way for practical, real-time video editing applications. Code is available at <https://github.com/KU-VGI/OSVE>.

1 Introduction

Text-guided video editing, *i.e.*, modifying a source video’s visual content according to a natural-language instruction, is increasingly vital for content creation, filmmaking, and interactive media. Recent generative diffusion models, spanning Text-to-Image (T2I) [10, 39, 44, 47] and Text-to-Video (T2V) [1, 8, 17, 34, 52] architectures, have enabled the generation of photorealistic, temporally coherent visual content from text alone. Building upon these powerful backbones, a growing body of work performs training-free, zero-shot video editing by intervening in the diffusion process [4, 7, 12, 14, 25, 53–55]. These methods typically follow a two-stage pipeline: (i) *inverting* the source video via multi-step inversion [30, 45] to recover its latent trajectory, and (ii) *generating* the edited frames by re-running the diffusion process under a modified text prompt, often with cross-frame attention control [9, 51] to preserve spatial structure and temporal layout.

* Corresponding author.



Fig. 1: Three frameworks for text-guided video editing. (a) *Multi-to-Multi*: The standard approach yields high-quality edits but is impractically slow. (b) *Multi-to-One*: Naïvely adopting a one-step diffusion model produces degraded, temporally inconsistent outputs. (c) *One-to-One (Ours)*: Our proposed framework uses a learnable one-step encoder to enable fast, temporally consistent, and high-quality video editing, resolving the speed-quality trade-off.

Despite the editing quality these methods achieve, they inherit a critical bottleneck: **extreme computational cost**. Because every frame must be processed over tens to hundreds of denoising steps, wall-clock time scales with both the number of frames and the number of steps. For instance, RAVE [14], reported as one of the fastest T2I-based editing methods, still requires approximately 24 hours to edit a 5-minute, 30FPS video¹. T2V backbones, which incorporate heavyweight temporal attention modules, demand even greater resources. This extreme latency renders current approaches impractical for real-time or streaming-style applications [56].

To overcome this bottleneck, we shift toward a *one-step* generation paradigm. One-step diffusion models [5, 27, 33, 59, 60], distilled from their multi-step teachers, have achieved competitive or even superior image quality, yet their potential for video editing remains unexplored. By adapting this ultra-fast regime to video, we build a One-to-One editing framework (one-step inversion followed by one-step generation) that replaces the costly iterative denoising loop with a single forward pass. We instantiate and validate our framework on a publicly released one-step T2I backbone (DMD2 [59]) to demonstrate its effectiveness with today’s most accessible one-step generators.

However, directly applying one-step diffusion models to video editing presents three critical challenges (see Figure 1(b)):

(i) Inversion-generation mismatch. Conventional multi-step inversion [45] is both *time-consuming* and fundamentally ill-suited for one-step generators. Since the inversion follows a multi-step trajectory while the generator performs a direct mapping, the mismatch between forward and reverse paths causes severe information loss, producing blurred and degraded results.

¹ Measured on a single NVIDIA RTX 6000 Ada GPU at 512×512 resolution.

(ii) **Structural collapse under single-step control.** Preserving the source video’s spatial structure (*e.g.*, a wolf’s rotating head) while altering its appearance (*e.g.*, to a bear) is critical for faithful editing. We empirically find that existing structure-preserving methods such as Prompt-to-Prompt [9] and ControlNet [61] fail when constrained to a single step. Unlike multi-step processes where control signals are distributed across steps and gradually absorbed, one-step models concentrate the same signal into a single pass with no room for correction, causing it to dominate rather than guide the output, a phenomenon we term “over-steering”.

(iii) **Temporal inconsistency.** One-step T2I diffusion models do not inherently encode temporal information, resulting in flickering or discontinuous frames when applied to video. Existing temporal consistency techniques, including attention fine-tuning [55], cross-frame attention control [4, 25, 36, 53], and noise shuffling across sampling steps [14], all rely on the multi-step structure and are therefore incompatible with the one-step regime.

To address these challenges, we propose **OSVE (One-Step Video Editing)**, a One-to-One video editing framework explicitly designed for one-step generation. For **inversion-generation mismatch**, the core of OSVE is a *learnable video encoder*, initialized from a pre-trained U-Net to inherit a strong prior over the generator’s latent space, that replaces slow multi-step inversion with a single forward pass. For **structural collapse**, we introduce a **Structure-Aware Editing (SAE)** loss that trains the encoder to produce latents inherently preserving structural integrity under modified text prompts, eliminating the need for external guidance that fails in one step. By training on diverse editing pairs, the encoder learns structure-aware latent representations that maintain the source’s layout while permitting appearance changes (Figure 1(c)). For **temporal inconsistency**, we propose **Unified-Frame Editing (UFE)**, which concatenates the inverted latents of consecutive frames into a single unified tensor, enabling the model’s self-attention to align features across frames in one pass. Combined with a sliding-window strategy for local coherence and anchor-frame selection for long-range stability, UFE ensures both short-term and long-term temporal consistency.

In summary, our main contributions are:

- We present **OSVE**, the first One-to-One video editing framework for the one-step diffusion regime, replacing multi-step inversion with a learnable single-pass encoder.
- We propose a **Structure-Aware Editing (SAE)** loss that directly injects structural preservation into the inversion stage, mitigating the over-steering problem inherent in single-step control.
- We introduce **Unified-Frame Editing (UFE)**, which enforces temporal consistency via unified latent processing with a sliding window and anchor frame, ensuring both short- and long-range coherence.
- Extensive experiments show that OSVE achieves **sub-second latency**, $\sim 155\text{--}171\times$ faster than the previous fastest method, while delivering comparable or superior visual quality, enabling practical real-time video editing.

2 Related Work

Text-Guided Video Editing with Diffusion Models. Building on advances in image-level generative editing, including GAN inversion [31, 32, 38, 50] and diffusion-based manipulation [2, 9, 21, 30], numerous zero-shot techniques have been proposed to extend such capabilities to video while preserving temporal consistency. These include fusing attention maps [35], propagating token features [7, 53], using optical flow guidance [4], and shuffling noise [14]. While these approaches have significantly improved video coherence, they all operate within a multi-step sampling framework. This iterative nature, where operations are repeated for each frame across numerous steps, results in prohibitively long run-times. This computational cost remains a critical barrier, motivating our work on high-quality, single-step video editing.

One-Step Diffusion Models. To overcome the significant latency of iterative diffusion sampling, a prominent line of research focuses on distilling multi-step teacher models into highly efficient one- or few-step generators. This paradigm was advanced by methods like Progressive Distillation and Consistency Models, with latent variants (LCM/LCM-LoRA) enabling 2–8 step generation for Stable Diffusion [29, 41, 46]. Subsequent works have further refined this approach. Rectified Flow and its successor, InstaFlow, straighten probability-flow trajectories to achieve SD-level quality in a single step via reflow-based distillation [26, 27]. Other strategies, such as Adversarial Diffusion Distillation (*e.g.*, SDXL-Turbo, SDXL-Lightning) and various distribution matching techniques (*e.g.*, DMD, SwiftBrush), couple teacher supervision with GAN losses or novel distillation objectives to achieve high-quality synthesis in just 1–4 steps [5, 22, 33, 42, 59, 60]. These one-step models have revolutionized image generation speed, but this progress has been exclusively image-centric. This creates a critical gap and a major opportunity, as the domain of video editing desperately needs such acceleration. Our work is the first to bridge this gap, introducing the necessary mechanisms to successfully adapt these fast generators for video, thereby unlocking their potential for practical, high-speed applications.

3 OSVE: One-Step Video Editing

We present OSVE, a framework for one-step video editing built on two core ideas: (i) encoding source structure into the initial noise latent so that a single-pass generator can preserve it without iterative control, and (ii) enforcing temporal consistency by letting all frames share context through a unified attention mechanism. Section 3.1 develops the first idea through a structure-aware inversion encoder trained with a dedicated editing loss. Section 3.2 introduces Unified-Frame Editing (UFE) for the second, along with sliding-window and anchor-based schemes that scale it to long videos. Figure 2 illustrates an overview of the framework.

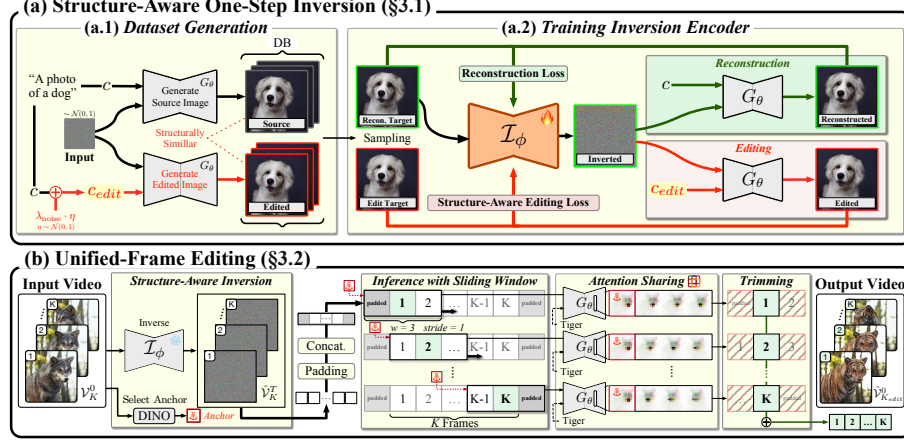


Fig. 2: Overview of our OSVE framework for one-step video editing. (a) *Training*: We train an inversion encoder using a novel dataset of structurally aligned image pairs generated via *Prompt Perturbation*. The encoder learns to predict an initial noise that supports both faithful reconstruction and structure-aware editing. (b) *Inference*: Our *Unified-Frame Editing (UFE)* method inverts all video frames, concatenates their latents into a single map, and processes them in one pass. This enables cross-frame attention, ensuring temporal consistency. An anchor-based sliding window scales the approach to long videos while maintaining global coherence.

3.1 Structure-Aware One-Step Inversion

Recent one-step diffusion models have dramatically accelerated image and video generation, yet their application to editing remains severely limited. The core difficulty is **structure preservation**: an edit should maintain the spatial layout, object pose, and geometric relationships of the source while modifying only the intended semantics. Structure preservation has been the central concern in diffusion-based editing from the outset, as methods such as Prompt-to-Prompt (P2P) [9], ControlNet [61], and PnP [51] were all designed primarily to keep the source structure intact during semantic modification. Without it, edits suffer from spatial misalignment and temporal inconsistency (Figure 1(b)), rendering them practically unusable regardless of semantic quality. While multi-step diffusion models have largely addressed this through iterative inversion and control, one-step models lack such mechanisms, making structure preservation the central bottleneck that must be solved before one-step editing becomes practical.

To understand why these existing solutions do not transfer, consider how they work in the multi-step setting. DDIM inversion [45] traces the denoising trajectory backward to recover a latent that approximately reconstructs the source, and the editing techniques listed above then intervene at selected steps along this trajectory. Crucially, these methods rely on the iterative nature of the process: each denoising step provides a corrective feedback loop that absorbs small perturbations introduced by the control signal, gradually steering the generation while keeping the structure intact. One-step models, however, collapse the entire

trajectory into a single forward pass, which creates two coupled difficulties. First, there is no multi-step trajectory to reverse, so standard DDIM-style inversion is inapplicable. Second, and more fundamentally, iterative control mechanisms lose the corrective loop they depend on. In iterative generation, the control signal is distributed across many denoising steps, and each step applies only a small perturbation that subsequent steps can absorb and correct. When the same control is concentrated into a single pass, its full influence is applied at once with no opportunity for subsequent correction, causing the control signal to dominate the generation rather than gently guide it. The resulting “over-steering” artifacts are shown in Figure 4(b, c, e, f).

This analysis points to a clear design principle: in a one-step setting, structure preservation cannot be imposed during generation but should be encoded before generation begins. If the initial noise latent itself already carries the structural information of the source frame, the one-step generator can preserve that structure in its single forward pass without any auxiliary control mechanism. We realize this principle by training an inversion encoder $\mathcal{I}_\phi(\cdot)$ that directly maps a source frame to an initial noise latent encoding the spatial structure needed for both faithful reconstruction and meaningful editing.

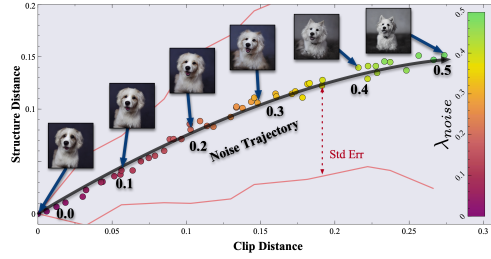
Concretely, given a VAE-encoded latent z^0 from a pre-trained autoencoder [16, 39] and a prompt embedding $c = E(y)$ for a given text prompt y (where E is the frozen text encoder, *e.g.*, CLIP [37]), our encoder predicts an initial noise latent $\hat{z}^T = \mathcal{I}_\phi(z^0, c)$, where T denotes the terminal diffusion timestep (*i.e.*, 999). Following previous inversion methods [9, 30, 38, 45, 50, 51], our objective is to predict a noise latent $\hat{z}^T$ that simultaneously satisfies two criteria: **reconstruction** and **editability**. To endow the encoder with structure-aware editability, we proceed in two stages: **1) Dataset generation with prompt perturbation**: We curate a dataset of structurally similar image pairs that share pose and layout but differ visually (Figure 2(a.1)). **2) Training with Structure-Aware Editing loss**: We train the encoder using a newly proposed structure-aware editing loss that simulates editing and constrains the output to align with the paired target (Figure 2(a.2)).

Dataset Generation with Prompt Perturbation. Training the inversion encoder to encode structure before generation requires data that teaches what “structure preservation under semantic change” looks like. Specifically, we need pairs of images that share spatial layout but differ in appearance, so that the encoder learns to produce latents capturing the shared structure while remaining agnostic to the varying semantics.

To generate such a dataset $DB = \{(z^0, z_{\text{edit}}^0, c, c_{\text{edit}})\}$, a natural approach would be to apply an off-the-shelf image editor to each source image. However, such editors typically rely on different generative backbones, and their outputs may fall outside the reachable manifold of the target one-step generator G_θ . Training $\mathcal{I}_\phi$ on these distributionally mismatched targets leads to inconsistent gradients, as the encoder is asked to predict latents for images that G_θ could never have produced in the first place. Table 1 confirms this: an encoder trained

Table 1: Quantitative comparison of reconstruction quality for different inversion strategies.

	PSNR $\uparrow$	LPIPS $\downarrow$	MSE $\downarrow$	SSIM $\uparrow$
DDIM [45]	13.87	0.525	0.045	0.446
Off-the-shelf ²	17.54	0.400	0.032	0.543
Ours	19.16	0.261	0.014	0.624

**Fig. 3:** Structural (DINO) and semantic (CLIP) distances as λ_{noise} varies.

on off-the-shelf² editing pairs yields notably lower reconstruction quality than ours across all metrics, indicating that the mismatched supervision corrupts the encoder’s overall latent representation. We also report DDIM 50-step inversion [45] as a baseline, which performs worst due to the fundamental inversion–generation mismatch discussed in Section 1.

To avoid this mismatch, we propose a simple yet effective alternative termed *prompt perturbation*, which ensures that every target image is generator-aligned, *i.e.*, lies within G_θ ’s output distribution by construction. Concretely, we first sample a text prompt from a corpus (LAION [43], JourneyDB [48]) and a random noise vector $z^T \sim \mathcal{N}(0, I)$. Let $c = E(y)$ be the text embedding of the sampled prompt y . We then construct a perturbed embedding as:

$$c_{\text{edit}} = c + \lambda_{\text{noise}} \cdot \eta, \quad \eta \sim \mathcal{N}(0, I), \quad (1)$$

where λ_{noise} controls the perturbation intensity. Because both the source image $G_\theta(z^T, c)$ and the edited image $G_\theta(z^T, c_{\text{edit}})$ are generated by the same frozen generator from the same initial noise, the resulting pairs are sampled from G_θ ’s reachable output distribution and tend to share a similar structural layout while primarily differing in appearance. This generator-aligned supervision provides a well-posed training signal for $\mathcal{I}_\phi$, which must predict latents that G_θ can faithfully decode.

We empirically study the effect of λ_{noise} on both structure and semantics (Figure 3). Starting from the base prompt “a photo of a dog”, we vary $\lambda_{\text{noise}} \in \{0.0, 0.1, \dots, 0.5\}$ and, for each setting, generate 100 image pairs using the same initial noise z^T . For each pair, we measure (i) a structural distance using DINO [3], and (ii) a semantic distance using CLIP [37]. As λ_{noise} increases, the semantic distance grows, while for appropriately small noise (*e.g.*, $\lambda_{\text{noise}} = 0.1$) the structural distance remains low and stable, confirming that small perturbations preserve layout while inducing meaningful appearance changes. Guided by this analysis, we set $\lambda_{\text{noise}} = 0.1$ in all experiments (see Supplementary B and G for training results of other parameters).

² We use Google Gemini’s Nano Banana as a representative off-the-shelf image editor with default recommended settings.

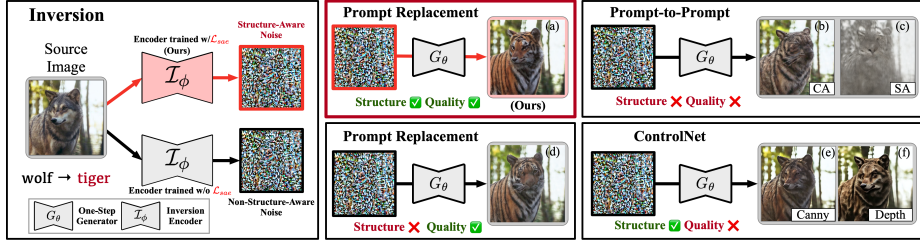


Fig. 4: Comparison of structure-preserving methods for one-step editing. (a) Our method, trained with the proposed $\mathcal{L}_{\text{sae}}$, maintains structural integrity. (d) Without $\mathcal{L}_{\text{sae}}$, our model fails, leading to structural collapse. (b, c, e, f) Similarly, existing control methods often produce degraded results, demonstrating their unsuitability for one-step generation. SA and CA denote the replacement of self-attention and cross-attention, respectively.

Given (c, c_{edit}) and a shared z^T , we obtain the reconstruction and editing targets as

$$\underbrace{z^0 = G_\theta(z^T, c)}_{\text{Reconstruction Target}}, \quad \underbrace{z_{\text{edit}}^0 = G_\theta(z^T, c_{\text{edit}})}_{\text{Editing Target}}. \quad (2)$$

This process yields a large dataset in which most pairs exhibit strong structural similarity but differ in appearance, effectively simulating the desired editing behavior.

Training with Structure-Aware Editing Loss. The dataset provides pairs that share structure but differ in semantics, each pair generated from the same noise z^T under the original and perturbed prompts. We translate this into a training objective that forces the inversion encoder to produce a latent $\hat{z}^T = \mathcal{I}_\phi(z^0, c)$ serving both reconstruction and structure-preserving editing. Our loss is:

$$\mathcal{L}_{\text{total}} = \underbrace{\|z^0 - G_\theta(\mathcal{I}_\phi(z^0, c), c)\|_2^2}_{\mathcal{L}_{\text{mse}}} + \lambda_{\text{sae}} \underbrace{\|z_{\text{edit}}^0 - G_\theta(\mathcal{I}_\phi(z^0, c), c_{\text{edit}})\|_2^2}_{\mathcal{L}_{\text{sae}}}, \quad (3)$$

where λ_{sae} is the weight parameter (set to 1.0), $z^0 = G_\theta(z^T, c)$, and $z_{\text{edit}}^0 = G_\theta(z^T, c_{\text{edit}})$.

The first term, $\mathcal{L}_{\text{mse}}$, ensures that the inverted latent allows G_θ to accurately reconstruct the original image z^0 . Reconstruction alone, however, does not guarantee that the same latent will preserve structure when the prompt changes. The second term, the Structure-Aware Editing (SAE) loss $\mathcal{L}_{\text{sae}}$, directly addresses this: it requires that the same inverted latent $\hat{z}^T$, when decoded with the perturbed prompt c_{edit} , produces an output matching the pre-generated target z_{edit}^0 . Because z_{edit}^0 was generated from the same noise z^T as z^0 , it shares the source structure by construction, so matching it teaches the encoder to produce latents whose structural content is invariant to prompt changes. Figure 4

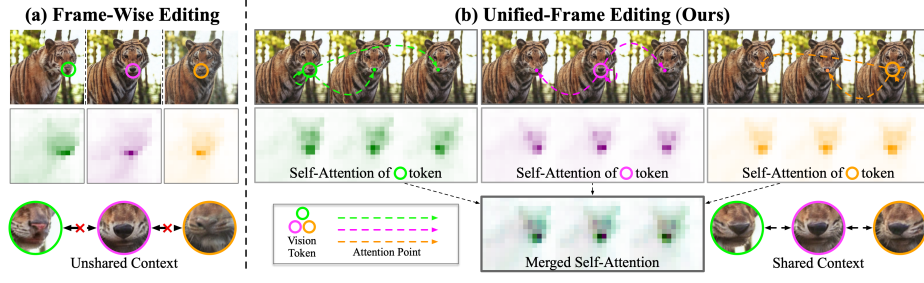


Fig. 5: Mechanism of Unified-Frame Editing for Temporal Consistency. (a) Frame-Wise Editing: Processing frames individually restricts attention to within each frame, preventing information sharing and causing temporal incoherence. (b) Unified-Frame Editing (UFE): Our method concatenates all frame latents, allowing the attention mechanism to operate globally. This enables features to be matched and aligned across all frames, enforcing temporal consistency throughout the edit.

validates this design: without $\mathcal{L}_{\text{sae}}$, the model suffers from structural collapse (d), while existing control methods such as attention injection (b, c, e, f) similarly fail, confirming that structure preservation must come from the latent itself rather than from external control during generation.

For implementation, we initialize the inversion encoder $\mathcal{I}_\phi$ with the pre-trained one-step model G_θ , leveraging components from its U-Net. Detailed training settings and visual and statistical analyses of the latents $\hat{z}^T$ produced by $\mathcal{I}_\phi$ are provided in Supplementary C.

3.2 Unified-Frame Editing

While the inversion encoder from Section 3.1 is effective for single images, applying it frame-wise to video does not preserve temporal consistency. Most prior video editing methods enforce consistency by injecting guidance signals into intermediate features or attention maps during multi-step denoising [4, 7, 14, 25, 53, 55, 57]. However, as discussed in Section 3.1, such iterative guidance mechanisms are inapplicable in the one-step regime (Figure 4).

This motivates a fundamentally different strategy: rather than manipulating internal attention, we encourage cross-frame communication by modifying only the *input* latent representation. We propose **Unified-Frame Editing (UFE)**, an input-level latent unification scheme in which inverted latents from multiple frames are spatially concatenated and processed simultaneously in a single forward pass (Figure 2(b)).

Concretely, given a video sequence of K frames $\mathcal{V}_K^0 = (z_0^0, \dots, z_{K-1}^0)$, each frame is independently inverted via our encoder:

$$\hat{z}_k^T = \mathcal{I}_\phi(z_k^0, c), \quad k = 0, \dots, K-1. \quad (4)$$

We denote the set of inverted latents as $\hat{\mathcal{V}}_K^T = (\hat{z}_0^T, \dots, \hat{z}_{K-1}^T)$. Let $\hat{z}_k^T \in \mathbb{R}^{C \times H \times W}$ denote the inverted latent for the k -th frame. The unified latent map Z_{UFE}^T is

formed by concatenating all K latents along the width dimension:

$$Z_{\text{UFE}}^T = \text{Concat}(\hat{z}_0^T, \hat{z}_1^T, \dots, \hat{z}_{K-1}^T, \text{axis} = W). \quad (5)$$

The resulting map Z_{UFE}^T has dimensions $\mathbb{R}^{C \times H \times (W \cdot K)}$ and is passed to the one-step generator $G_\theta(\cdot)$ with the editing prompt c_{edit} . This is feasible because the U-Net operates on spatial features and naturally accepts inputs of arbitrary size [28, 39, 40]. Crucially, this design allows the generator’s self-attention to function globally across all frames, leveraging its ability to align corresponding features [49] and thereby enforce temporal consistency without any external guidance module. Figure 5 visualizes this effect: unlike frame-wise editing where information remains isolated, a token representing a specific feature (*e.g.*, the tiger’s nose) in one frame can directly attend to its counterpart in other frames, ensuring that edits are applied consistently across the entire sequence.

Sliding Window with Anchor Frame. While UFE is effective, its VRAM requirement and inference time scale with the number of frames (Supplementary H), making it computationally challenging for long videos. To address this, we extend UFE with a sliding window mechanism guided by a global anchor frame. Figure 2(b) illustrates the pipeline of our method. The core idea is to process the video in manageable chunks while ensuring both local and global temporal consistency. The sliding window provides local coherence, while the anchor frame enforces a consistent global style and content reference across all windows. First, we identify a representative anchor frame by selecting the medoid in the DINO [3] feature space. Its inverted latent, $\hat{z}_A^T$, serves as a global anchor. We validate this choice against alternative strategies, including the use of multiple anchors for videos with scene transitions or dynamic content changes, in Supplementary D.

We then process the video using an overlapping sliding window approach. For each window of w inverted latents, we prepend the anchor latent $\hat{z}_A^T$ before feeding the concatenated map to the generator G_θ . After generation, the output corresponding to the anchor is discarded, and we extract only the central s frames from the window’s output. These segments are then seamlessly stitched together to form the final edited video. This anchor-based sliding window strategy effectively prevents temporal drift by ensuring that each segment is coherent both locally and globally. We set the window size as $w = 7$ and a stride of $s = 5$ through ablation (Supplementary E). A detailed, step-by-step formulation of this process is provided in Supplementary F.

4 Experiments

4.1 Experimental Setup

We evaluate our framework across three video editing settings. The first, **Multi-to-Multi**, uses multi-step diffusion models for both inversion and editing. We compare against five baselines: FLATTEN [4], TokenFlow [36], FRESCO [57], RAVE [14], and COVE [53], all built on Stable Diffusion 1.5 [39]. The second, **Multi-to-One**, performs video inversion via multi-step DDIM while editing

Table 2: Performance comparison on short (20 frames) videos.

Framework	Method	SC	BC	TF	MS	AQ	IQ	BQS	FPS
Multi-to-Multi	FLATTEN [‡]	0.965	0.970	0.964	0.972	0.625	0.639	0.611	0.072
	TokenFlow [‡]	0.983	0.976	0.985	0.991	0.668	0.680	0.663	0.075
	FRESCO [‡]	0.978	0.974	0.973	0.991	0.649	0.729	0.674	0.078
	RAVE [‡]	0.982	0.976	0.975	0.986	0.637	0.695	0.653	0.091
	COVE [‡]	0.983	0.976	0.984	0.989	0.645	0.655	0.639	0.061
Multi-to-One	Prompt Replacement [†]	0.921	0.946	0.948	0.979	0.593	0.562	0.548	4.732
	Prompt-to-Prompt [†]	0.915	0.945	0.965	0.978	0.587	0.566	0.548	0.898
	ControlNet (Depth) [†]	0.968	0.961	0.970	0.984	0.658	0.673	0.646	0.578
	ControlNet (Canny) [†]	0.960	0.969	0.966	0.983	0.582	0.662	0.603	0.578
	Plug-and-Play [†]	0.953	0.971	0.995	0.995	0.486	0.220	0.345	0.398
One-to-One	OSVE (Ours)[†]	0.983	0.977	0.978	0.991	0.678	0.703	0.678	15.625

[‡] Multi-Step Diffusion Model (SD1.5). [†] One-Step Diffusion Model (DMD2).

Table 3: Performance comparison on long (90 frames) videos.

Framework	Method	SC	BC	TF	MS	AQ	IQ	BQS	FPS
Multi-to-Multi	FLATTEN [‡]	0.933	0.960	0.965	0.975	0.648	0.636	0.615	0.052
	TokenFlow [‡]	0.974	0.975	0.986	0.990	0.622	0.668	0.633	0.084
	FRESCO [‡]	0.956	0.968	0.987	0.992	0.640	0.702	0.655	0.087
	RAVE [‡]	0.962	0.963	0.978	0.983	0.665	0.692	0.659	0.102
	COVE [‡]	0.955	0.964	0.985	0.988	0.651	0.650	0.633	0.041
Multi-to-One	Prompt Replacement [†]	0.841	0.912	0.970	0.975	0.561	0.528	0.503	4.732
	Prompt-to-Prompt [†]	0.900	0.956	0.969	0.974	0.423	0.528	0.451	0.898
	ControlNet (Depth) [†]	0.945	0.944	0.971	0.978	0.626	0.652	0.613	0.578
	ControlNet (Canny) [†]	0.934	0.952	0.967	0.976	0.590	0.660	0.598	0.578
	Plug-and-Play [†]	0.841	0.912	0.994	0.995	0.561	0.267	0.387	0.398
One-to-One	OSVE (Ours)[†]	0.958	0.965	0.978	0.989	0.670	0.723	0.677	15.793

[‡] Multi-Step Diffusion Model (SD1.5). [†] One-Step Diffusion Model (DMD2).

in a single step. As no prior work exists for this configuration, we construct baselines by adapting image editing techniques: Prompt Replacement, Prompt-to-Prompt [9], ControlNet [61] with depth and Canny guidance, and Plug-and-Play [51]. The third is our proposed **One-to-One** framework, which employs the pre-trained encoder for single-step inversion and edits via Prompt Replacement. For the Multi-to-One and One-to-One settings, we use DMD2 [59] as the one-step diffusion backbone. Further details are provided in Supplementary A.

For evaluation, we curated 60 videos from open platforms such as Pixabay and prior works, comprising 51 short videos (20 frames) and 9 long videos (90 frames). We designed five editing prompts per video—three for local edits (*e.g.*, object change, deletion) and two for global edits (*e.g.*, style transfer, background modification)—resulting in 255 short-video and 45 long-video evaluation pairs. Short videos are processed as a single window, while long videos use a sliding window of 7 frames with a stride of 5. All experiments were conducted on a single NVIDIA RTX 6000 Ada GPU.

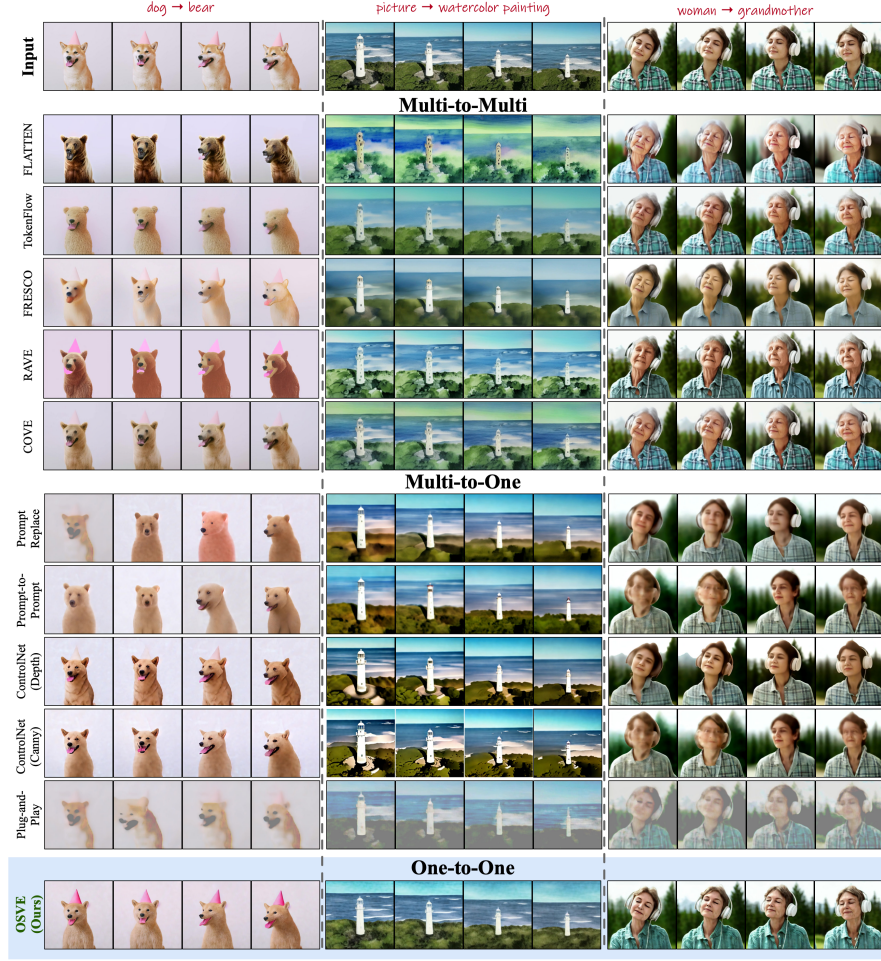


Fig. 6: Qualitative comparison with baseline methods.

4.2 Quantitative Comparison

We adopt VBench [11] metrics for quantitative evaluation, organized into two categories. **Temporal consistency** metrics include: Subject Consistency (SC), measuring subject identity preservation via DINO [3] features; Background Consistency (BC), assessing background stability via CLIP [37] features; Temporal Flickering (TF), where higher indicates less flickers; and Motion Smoothness (MS), evaluating motion fluidity using a pre-trained video interpolation model [20]. **Per-frame quality** metrics include: Aesthetic Quality (AQ), scored by the LAION aesthetic predictor [18], and Imaging Quality (IQ), quantifying perceptual artifacts using MUSIQ [15]. We also introduce the Balanced Quality

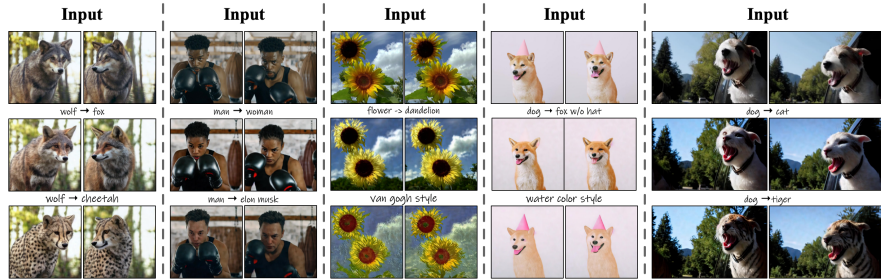


Fig. 7: Qualitative results with OSVE across various editing tasks.

Score (BQS): $BQS = \left(\frac{SC+BC+TF+MS}{4} \right) \times \left(\frac{AQ+IQ}{2} \right)$, which balances both dimensions multiplicatively, penalizing methods that excel in only one.

Tables 2 and 3 present the results for short and long videos, respectively. Our method, OSVE, achieves the highest BQS in both settings, demonstrating strong temporal consistency and per-frame quality simultaneously. In terms of efficiency, OSVE is approximately $171\times$ faster for short videos and $155\times$ faster for long videos compared to RAVE, the fastest Multi-to-Multi baseline. See Supplementary A for the speed comparison setup.

4.3 Qualitative Comparison

Figure 6 provides a qualitative comparison. OSVE successfully modifies objects while preserving their underlying structure, whereas the Multi-to-One baselines (rows 7–11) frequently exhibit structural collapse or visual degradation. Figure 7 illustrates the versatility of OSVE across diverse editing tasks, including object replacement (columns 1, 3, and 5), human identity modification (column 2), artistic style transfer (columns 3 and 4), and prompt-guided object removal (column 4).

4.4 User Study

We conducted a user study with 26 participants. Each participant viewed six editing methods applied to three randomly sampled videos from our datasets and rated each result on Temporal Consistency (TC) and Visual Quality (VQ) using a 5-point Likert scale. As shown in Table 4, OSVE achieves the highest scores on both dimensions, outperforming all baselines.

4.5 Ablation Study

Inversion Encoder. Table 5 validates the Structure-Aware Editing (SAE) Loss on the PIE-Bench [13], using structure distance [51] and CLIP scores [37]. While SAE Loss alone improves structure preservation, combining it with the reconstruction loss ($\mathcal{L}_{mse}$) yields the best performance in both structural integrity and editability, indicating a synergistic effect between the two objectives.

Table 4: User study results. TC: Temporal Consistency, VQ: Visual Quality (higher is better).

Method	TC	VQ
FLATTEN	2.59	2.56
TokenFlow	3.68	3.08
FRESCO	3.02	2.76
RAVE	2.74	2.70
COVE	3.59	3.41
OSVE (Ours)	3.85	3.65

Table 5: Ablation study on encoder training. We evaluate the effect of each loss component on PIE-Bench.

$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{sae}}$	Struct. Dist. ($\downarrow$)	CLIP Whole ($\uparrow$)	CLIP Edit ($\uparrow$)
✓	–	0.087	21.797	19.884
–	✓	0.074	22.349	19.863
✓	✓	0.064	22.329	20.416

Table 6: Ablation study of UFE on 90-frame videos.

Sliding window	Anchor	SC	BC	TF	MS	AQ	IQ	BQS
–	–	0.931	0.948	0.972	0.980	0.680	0.722	0.671
✓	–	0.954	0.963	0.977	0.988	0.671	0.723	0.676
–	✓	0.943	0.950	0.976	0.985	0.672	0.722	0.669
✓	✓	0.958	0.965	0.978	0.989	0.670	0.723	0.677

Unified Frame Editing. Table 6 presents the ablation study of Unified-Frame Editing (UFE). Both the sliding window and anchor-frame strategies individually improve temporal consistency (SC, BC, TF, MS), and their combination yields the strongest results. We observe a marginal decrease in AQ when applying UFE, likely due to processing the expanded latent representation. However, this minor trade-off is perceptually negligible compared to the substantial reduction in flickering and inconsistencies.

5 Discussion

We build OSVE on a one-step T2I backbone because current one-step T2V generators, despite recent progress in T2V distillation [23, 58, 63], still suffer from severe temporal artifacts and blurriness that make them unsuitable as a reliable editing backbone (see Supplementary G). In contrast, mature one-step T2I backbones have been reported to even surpass their multi-step teachers in generation quality [59], providing a clean testbed to evaluate our core contributions independently of backbone limitations.

The generative landscape is rapidly expanding toward T2V, T2AV, and other multimodal architectures, and one-step distillation of these models is an active area of research [6, 19, 23, 24, 62]. As their fidelity improves, extending one-step editing to these backbones will become both feasible and important. We believe the core principles behind OSVE, namely encoding structure into the initial latent and enforcing cross-frame consistency at the input level, are not specific to T2I and can generalize across modalities.

6 Conclusion

We presented **OSVE**, a framework that advances text-guided video editing by introducing a novel One-to-One pipeline. OSVE unlocks the potential of one-step diffusion models for this task, overcoming the speed limitations of conventional multi-step methods. Our two core contributions, a fast learnable inversion encoder trained with **SAE** loss to preserve structure and the **UFE** mechanism to enforce temporal coherence, resolve the key bottlenecks that previously hindered this approach. Our experiments show that OSVE achieves competitive quality and consistency at a fraction of the computational cost, making real-time video editing practical and accessible.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25480253), and in part by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korea government (MOTIE) (RS-2024-00444344), and by the “Advanced GPU Utilization Support Program” funded by the Government of the Republic of Korea (Ministry of Science and ICT).

References

1. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., Li, Y., Rubinstein, M., Michaeli, T., Wang, O., Sun, D., Dekel, T., Mosseri, I.: Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945 (2024). <https://doi.org/10.48550/arXiv.2401.12945>
2. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800 (2022)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
4. Cong, Y., Xu, M., Simon, C., Chen, S., Ren, J., Xie, Y., Perez-Rua, J.M., Rosenhahn, B., Xiang, T., He, S.: Flatten: optical flow-guided attention for consistent text-to-video editing (2024), <https://arxiv.org/abs/2310.05922>
5. Dao, T., Nguyen, T.H., Le, T., Vu, D., Nguyen, K., Pham, C., Tran, A.: Swift-brush v2: Make your one-step diffusion model better than its teacher. In: European Conference on Computer Vision. pp. 176–192. Springer (2024)
6. Ding, Z., Jin, C., Liu, D., Zheng, H., Singh, K.K., Zhang, Q., Kang, Y., Lin, Z., Liu, Y.: Dollar: Few-step video generation via distillation and latent reward optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
7. Geyer, M., Bar-Tal, O., Bagon, S., Dekel, T.: Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373 (2023)

8. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion (2024), <https://arxiv.org/abs/2501.00103>
9. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models pp. 6840–6851 (2020)
11. 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)
12. Jiang, Z., et al.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025). <https://doi.org/10.48550/arXiv.2503.07598>
13. Ju, X., Zeng, A., Bian, Y., Liu, S., Xu, Q.: Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=FoMZ4ljhVw>
14. Kara, O., Kurtkaya, B., Yesiltepe, H., Reh, J.M., Yanardag, P.: Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6507–6516 (2024)
15. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5148–5157 (2021)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)
17. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024). <https://doi.org/10.48550/arXiv.2412.03603>
18. LAION-AI: aesthetic-predictor. <https://github.com/LAION-AI/aesthetic-predictor> (2022), accessed: 2025-09-25
19. Li, J., Feng, W., Fu, T.J., Wang, X., Basu, S., Chen, W., Wang, W.Y.: T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. In: Advances in Neural Information Processing Systems (2024)
20. Li, Z., Zhu, Z.L., Han, L.H., Hou, Q., Guo, C.L., Cheng, M.M.: Amt: All-pairs multi-field transforms for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9801–9810 (2023)
21. Lim, H., Won, Y., Seo, J., Park, G.M.: Conceptsplitt: Decoupled multi-concept personalization of diffusion models via token-wise adaptation and attention disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
22. Lin, S., Wang, A., Yang, X.: Sdxl-lightning: Progressive adversarial diffusion distillation. arXiv preprint arXiv:2402.13929 (2024)

23. 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)
24. Lin, S., Yang, X.: Animatediff-lightning: Cross-model diffusion distillation. arXiv preprint arXiv:2403.12706 (2024)
25. Liu, S., Zhang, Y., Li, W., Lin, Z., Jia, J.: Video-p2p: Video editing with cross-attention control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8599–8608 (2024)
26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
27. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: Instaflo: One step is enough for high-quality diffusion-based text-to-image generation. In: The Twelfth International Conference on Learning Representations (2023)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
29. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
30. 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)
31. Moon, S.J., Kim, C., Park, G.M.: WINE: Wavelet-guided GAN inversion and editing for high-fidelity refinement. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
32. Moon, S.J., Park, G.M.: Intereststyle: Encoding an interest region for robust stylegan inversion. In: European Conference on Computer Vision (ECCV). pp. 460–476. Springer (2022)
33. Nguyen, T.H., Tran, A.: Swiftbrush: One-step text-to-image diffusion model with variational score distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7807–7816 (2024)
34. OpenAI: Video generation models as world simulators. Tech. rep., OpenAI (Feb 2024)
35. Qi, C., Cun, X., Zhang, Y., Lei, C., Wang, X., Shan, Y., Chen, Q.: Fatezero: Fusing attentions for zero-shot text-based video editing (2023)
36. Qu, L., Zhang, H., Liu, Y., Wang, X., Jiang, Y., Gao, Y., Ye, H., Du, D.K., Yuan, Z., Wu, X.: Tokenflow: Unified image tokenizer for multimodal understanding and generation (2024), <https://arxiv.org/abs/2412.03069>
37. 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)
38. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. pp. 10684–10695 (2022)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

41. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
42. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision. pp. 87–103. Springer (2024)
43. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
44. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (07–09 Jul 2015), <https://proceedings.mlr.press/v37/sohl-dickstein15.html>
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
46. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
47. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
48. Sun, K., Pan, J., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., et al.: Journeydb: A benchmark for generative image understanding. Advances in neural information processing systems **36**, 49659–49678 (2023)
49. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. Advances in Neural Information Processing Systems **36**, 1363–1389 (2023)
50. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
51. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1921–1930 (2023)
52. 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>
53. 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)
54. Wang, Y., et al.: Videodirector: Precise video editing via text-to-video models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
55. Wu, J.Z., Ge, Y., Wang, X., Lei, 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 (2023), <https://arxiv.org/abs/2212.11565>
56. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models. ACM Computing Surveys **57**(2), 1–42 (2024)

57. Yang, S., Zhou, Y., Liu, Z., Loy, C.C.: Fresco: Spatial-temporal correspondence for zero-shot video translation (2024), <https://arxiv.org/abs/2403.12962>
58. Yang, Y., Li, X., Zhang, H., Li, Z., Shechtman, E., Li, Y.: Towards one-step causal video generation via adversarial self-distillation. arXiv preprint arXiv:2511.01419 (2025)
59. 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)
60. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6613–6623 (2024)
61. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023)
62. Zhang, Z., Li, Y., Wu, Y., Xu, Y., Kag, A., Skorokhodov, I., Menapace, W., Siarohin, A., Cao, J., Metaxas, D., Tulyakov, S., Ren, J.: Sf-v: Single forward video generation model. In: *Advances in Neural Information Processing Systems* (2024)
63. Zheng, K., Wang, Y., Ma, Q., Chen, H., Zhang, J., Balaji, Y., Chen, J., Liu, M.Y., Zhu, J., Zhang, Q.: Large scale diffusion distillation via score-regularized continuous-time consistency. arXiv preprint arXiv:2510.08431 (2025)