ContextFlow: Training-Free Video Object Editing via Adaptive Context Enrichment

Yiyang Chen[1], Xuanhua He[2*], Xiujun Ma[1*], Jack Ma[2*]

[1]State Key Laboratory of General Artifical Intelligence, Peking University, Beijing, China
[2]The Hong Kong University of Science and Technology
chenyy@stu.pku.edu.cn, maxiujun@pku.edu.cn, {xhecd, ymacn}@connect.ust.hk

Abstract

Training-free video object editing aims to achieve precise object-level manipulation, including object insertion, swapping, and deletion. However, it faces significant challenges in maintaining fidelity and temporal consistency. Existing methods, often designed for U-Net architectures, suffer from two primary limitations: inaccurate inversion due to first-order solvers, and contextual conflicts caused by crude "hard" feature replacement. These issues are more challenging in Diffusion Transformers (DiTs), where the unsuitability of prior layer-selection heuristics makes effective guidance challenging. To address these limitations, we introduce ContextFlow, a novel training-free framework for DiT-based video object editing. In detail, we first employ a high-order Rectified Flow solver to establish a robust editing foundation. The core of our framework is Adaptive Context Enrichment (for specifying *what* to edit), a mechanism that addresses contextual conflicts. Instead of replacing features, it enriches the self-attention context by concatenating Key-Value pairs from parallel reconstruction and editing paths, empowering the model to dynamically fuse information. Additionally, to determine where to apply this enrichment (for specifying *where* to edit), we propose a systematic, data-driven analysis to identify task-specific vital layers. Based on a novel Guidance Responsiveness Metric, our method pinpoints the most influential DiT blocks for different tasks (*e.g.*, insertion, swapping), enabling targeted and highly effective guidance. Extensive experiments show that ContextFlow significantly outperforms existing training-free methods and even surpasses several state-of-the-art training-based approaches, delivering temporally coherent, high-fidelity results.

Page — https://yychen233.github.io/ContextFlow-page
Extended version — https://arxiv.org/abs/2509.17818

1 Introduction

Video object editing aims to achieve a range of challenging object-related editing tasks, including object insertion, swapping, and deletion. Unlike original video editing, video object editing requires the model to meticulously preserve the unmodified background while seamlessly integrating the edited object into the video's original context. This is a task that demands high spatial and temporal consistency.

Currently, there are two primary technical paths in the research community: training-based and training-free methods. (1) Training-based methods aim to build powerful and feed-forward models. Recent examples include video propagation-based models like I2V Edit (Ouyang et al. 2024), GenProp (Liu et al. 2025a) and ReVideo (Mou et al. 2024), as well as other architectures such as VideoAnyDoor (Tu et al. 2025), GetIn (Zhuang et al. 2025), VACE (Jiang et al. 2025) and UNIC (Ye et al. 2025), which achieve impressive results on respective benchmarks. However, these training-based methods are limited by the prohibitive computational costs and the demand for expensive large-scale datasets. (2) Compared with the training-based methods, the training-free method offers a more flexible and cost-effective alternative. Early works, like AnyV2V (Ku et al. 2024), leverage the vast knowledge embedded in pre-trained foundation models, obviating the need for any task-specific fine-tuning. This paradigm typically relies on a foundational workflow: first, inverting the source video to a noise latent using DDIM Inversion (Song, Meng, and Ermon 2021), and then guiding the new generation via Plug-and-Play (PnP) feature injection (Graikos et al. 2022). In this process, key internal features from a reconstruction of the source video are injected into the generation process of the edited video to enforce structural consistency.

However, this workflow faces limitations. It struggles with fidelity, leading to artifacts, inconsistent object identity, and difficulty in preserving the original background. These issues are further amplified by the recent architectural shift from U-Nets to Diffusion Transformers (DiTs), as traditional guidance mechanisms are ill-suited for the new model class. We provide a detailed analysis in Section 3.

To address these critical limitations, we propose ContextFlow, a novel training-free framework that significantly advances the editing process for DiT-based models. Instead of relying on lossy inversion and crude feature replacement, ContextFlow establishes a high-fidelity, highly reversible foundation for editing based on RF-Solver (Wang et al. 2025b). At its core is a novel Adaptive Context Enrichment mechanism, which empowers the model to dynamically fuse information from the original video and the desired edit on a per-token basis. This dynamic approach addresses the con-

*Corresponding authors.

Figure 1: Showcase of ContextFlow. ContextFlow achieves versatile and high-fidelity video object editing without any training, and demonstrates superior ability in many object-related challenging tasks, including object insertion, swapping, and deletion.

flict between content preservation and synthesis. To apply this guidance with efficiency and precision, we propose a systematic, data-driven Vital Layer Analysis to identify the most crucial intervention points within the DiT architecture.

Our contributions can be summarized as follows:

- We propose ContextFlow, a novel training-free framework that is the first to apply Rectified Flow inversion to video object editing. This establishes a high-fidelity foundation that significantly reduces editing artifacts.
- We design an adaptive context enrichment mechanism that addresses the contextual conflict of feature replacement. By concatenating Key-Value pairs, our method provides "soft guidance" that effectively balances the trade-off between edit fidelity and content preservation.
- We introduce a systematic and data-driven Vital Layer Analysis to identify the most crucial blocks for context injection in DiTs. This replaces the heuristic-based layer selection of U-Net frameworks and enables targeted, efficient guidance.
- Extensive experiments on diverse editing tasks, including object insertion, deletion, and swapping, demonstrate that ContextFlow significantly outperforms existing training-free approaches and even surpasses several state-of-the-art training-based methods.

2 Related Work

Diffusion-Based Video Editing The landscape of video editing has been reshaped by diffusion models. The field is broadly bifurcated into tuning-based and training-free methods. Tuning-based approaches, though powerful, often incur significant computational costs by adapting pre-trained models. This includes techniques like per-video optimization (Wu et al. 2023; Ouyang et al. 2024; Gao et al. 2025), subject-driven personalization (Ruiz et al. 2023; Molad et al. 2023; Wu et al. 2025; Wei et al. 2024), and content-motion decoupling for finer control (Mou et al. 2024; Tu et al. 2025). In contrast, training-free methods pursue zero-shot editing. Early efforts focused on preserving structure by manipulating cross-frame attention (Geyer et al. 2024; Qi et al. 2023). A dominant recent paradigm, however, propagates edits from a modified initial frame, leveraging strong I2V priors for impressive consistency (Ku et al. 2024; Bai et al. 2024). Concurrently, other works enhance motion and appearance control by customizing the diffusion process itself (Jeong, Park, and Ye 2024; Kara et al. 2024; Wang et al. 2025a; Burgert et al. 2025).

Figure 2: Motivation for ContextFlow. We highlight two core failures of prior methods: DDIM inversion causes poor reconstruction results (Figure 2c) compared to original frames (Figure 2a), while "hard replacement" leads to misaligned attentions (Figure 2d) that fail to focus on the intended edits (Figure 2b).

Reference-Guided Object Editing To overcome the ambiguity of text-only guidance, reference-guided editing uses an image to define an object's visual identity. A common principle is establishing feature-level correspondence to guide attention and ensure faithful appearance propagation (Su et al. 2025; Gu et al. 2024; Ku et al. 2024; Ouyang et al. 2024; Gao et al. 2025). A more formidable challenge is inserting a new object into a video. Novel methods tackle this by ensuring identity preservation, temporal coherence, and plausible scene interaction (Saini et al. 2024; Zhuang et al. 2025; Shen et al. 2025). Frameworks like VideoAnydoor (Tu et al. 2025) and MotionCtrl (Wang et al. 2024) provide a motion prompt-based object insertion scheme to control the motion of objects in detail. Our approach aims to unify object-related video editing tasks by constructing a training-free framework based on the latest DiT model.

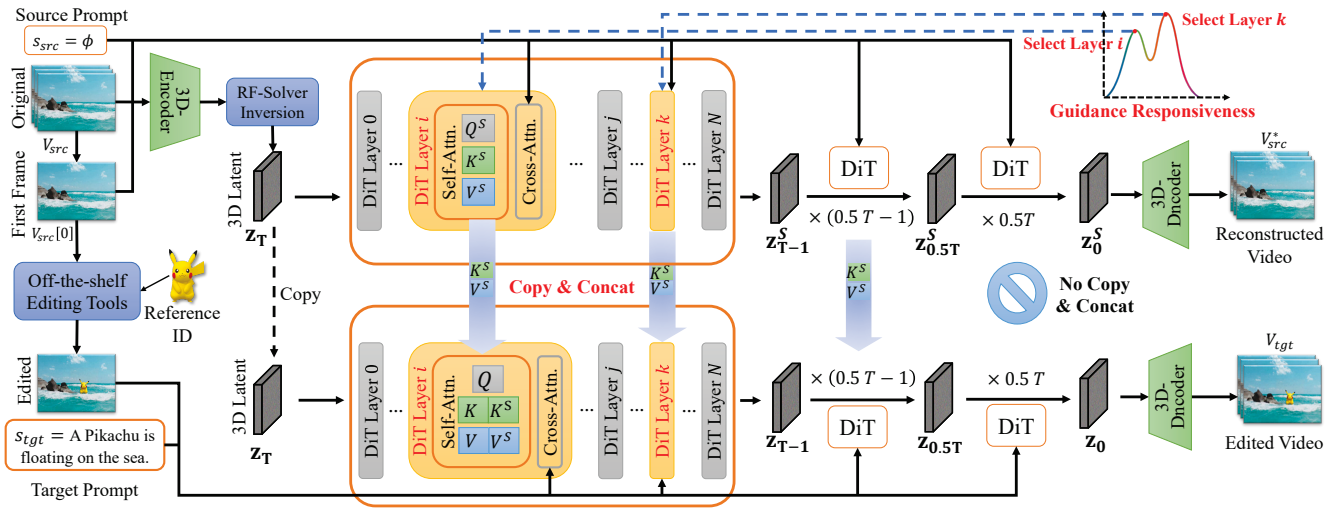

Figure 3: Overview of the ContextFlow. Our method begins with a high-fidelity video inversion using RF-Solver to obtain a shared noise latent $\mathbf{z}_T$. A dual-path sampling process then decouples reconstruction and editing. The editing path is guided by our core mechanism, Adaptive Context Enrichment, where Key-Value pairs from the reconstruction path are concatenated into the self-attention blocks of the editing path. This guidance is precisely targeted to vital layers, identified via our Guidance Responsiveness analysis, and is only active during the first half of the denoising process to balance fidelity and consistency.

3 Method

Given a reference video, we aim to edit objects via insertion, swapping, and deletion. In the following, we first analyze core challenges in training-free video object editing and our motivation, then present an overview of the proposed ContextFlow, followed by details of high-fidelity inversion, adaptive context enrichment, and Vital layer analysis.

Core Challenges and Motivations

The Challenge of Video Inversion for Editing For training-free editing, a critical first step is to invert a real video V back to its corresponding noise latent z_1. This process should ideally be perfectly reversible, creating an unambiguous anchor that encodes all spatiotemporal information of the source video. This is typically modeled as solving an Ordinary Differential Equation (ODE) that describes the path from noise to data. However, in practice, this ODE is solved numerically. Standard techniques like DDIM Inversion, which rely on first-order solvers similar to the Euler method, discretize the path from z_1 to z_v into N steps:

$$z_{t_{i-1}} = z_{t_i} + (t_{i-1} - t_i) v_\theta\left(z_{t_i}, t\right) \tag{1}$$

where t_i goes from 1 to 0. Naively reversing this process for inversion introduces significant discretization errors, which accumulate over timesteps. This results in a noisy latent that cannot faithfully reconstruct the original video, as shown in Figure 2, severely compromising the quality and consistency of any subsequent edits.

Contextual Conflict in Guidance The conventional PnP mechanism, which involves a "hard" replacement of features, is often too crude. This rigid intervention can create a conflict between the source video's structure and the edited object, leading to visual artifacts and inconsistent object identity. For example, queries from the editing path, seeking to form a new concept, are forced to attend to keys from the original video. This mismatch confuses the attention mechanism, leading to suppressed edits or artifacts. As visualized in Figure 2, hard replacement creates a semantic conflict, causing the edit-related queries to erroneously attend to keys and values from the original, irrelevant video context. This issue is compounded by the move to homogeneous Diffusion Transformers, whose lack of distinct semantic layers—unlike hierarchical U-Nets—makes it unclear where and how to effectively inject guidance. In response to these three core challenges, we propose ContextFlow, a novel framework designed specifically for high-fidelity, DiT-based video object editing.

Overview of ContextFlow

Our proposed framework, ContextFlow, addresses this by designing a controlled generation process within a pre-trained I2V Diffusion Transformer, all without any weight modification. As illustrated in Figure 3, our approach is built upon three foundational parts. First, to solve the inversion fidelity problem, we establish a near-lossless and highly reversible foundation using Rectified Flow, creating a clean canvas for editing. Next, to address the contextual conflict inherent in feature replacement, we introduce an Adaptive Context Enrichment mechanism. Instead of a crude "hard" replacement, this method enriches the context by concatenating the Key-Value (KV) pairs from both the source video's reconstruction and the editing path, empowering the DiT's self-attention to dynamically balance preservation and synthesis. Finally, to answer the critical question of where to apply this guidance, our data-driven Vital Layer Analysis systematically identifies the most crucial layers for

intervention. This avoids the drawbacks of naive all-layer injection and replaces unreliable heuristics, ensuring both precision and efficiency. Together, these components enable robust, high-fidelity video editing in a training-free manner.

High-Fidelity Inversion via Rectified Flow

Our editing workflow begins with the source video V_{src} and a target prompt s_{tgt}. Using an off-the-shelf image editor (e.g., AnyDoor (Chen et al. 2024)), we first modify the initial frame $V_{src}[0]$ to create the edited frame I_{edit}. The primary challenge is then to propagate the static edit in I_{edit} throughout the video, guided by s_{tgt}, while maintaining fidelity to V_{src}.

As established in our preliminaries, the success of this editing hinges on the quality of the initial noise latent. Standard inversion methods like DDIM are lossy, making it difficult to disentangle editing artifacts from inversion errors. To eliminate this ambiguity, a higher-order numerical solver is essential. Therefore, we introduce RF-Solver (Wang et al. 2025b), a training-free, second-order sampler that provides the highly reversible and high-fidelity mapping required for a robust generative foundation. RF-Solver achieves its precision by utilizing a second-order Taylor expansion to more accurately estimate the ODE path during inversion:

$$z_{t_{i+1}} = z_{t_i} + (t_{i+1} - t_i) v_\theta(z_{t_i}, t_i) + \frac{1}{2}(t_{i+1} - t_i)^2 v_\theta^{(1)}(z_{t_i}, t_i) \tag{2}$$

where $v_\theta^{(1)}$ is the numerically estimated time derivative of the velocity field. We apply this to the VAE-encoded latents of the source video, conditioned on its original first frame $V_{src}[0]$ and a null-text prompt $s_{src} = \phi$, to obtain a unique noise anchor $\mathbf{z_1}$:

$$\mathbf{z_1} = \text{RF-Solver}_{\text{inversion}}(V_{src}, V_{src}[0], s_{src}) \tag{3}$$

The resulting anchor $\mathbf{z_1}$ provides a faithful representation of the original video's spatiotemporal information, establishing a robust foundation for our editing mechanism

Adaptive Context Enrichment

With a reliable noise anchor $\mathbf{z_1}$ established, our core challenge becomes propagating the edit from the triplet $(\mathbf{z_1}, I_{edit}, s_{tgt})$ while preserving the original video's structure. A naive, single-path generation, denoising from $\mathbf{z_1}$ using only the new conditions (I_{edit}, s_{tgt}) is insufficient as it lacks continuous guidance from the source video, leading to content drift. While prior methods use feature injection from the original source video, their reliance on "hard replacement" creates a fundamental problem in DiTs.

Contextual Conflicts of Hard Replacement We identify the core limitation of naive feature injection as a contextual conflict. This occurs when edit-specific queries (Q^{edit}), which seek to form a new semantic concept (e.g., "a Pikachu is floating on the sea."), are forced to attend to a context (K^{res}, V^{res}) from the original video that only contains information about the original scene (e.g., "the sea surface"). As illustrated in Figure 4, this semantic mismatch confuses the attention mechanism, leading to suppressed edits or artifacts. This necessitates a more intelligent fusion strategy.

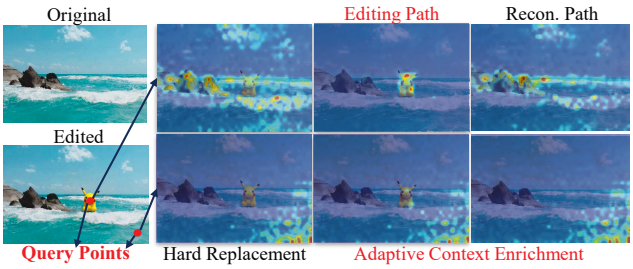

Figure 4: Resolving Contextual Conflict. Hard replacement misdirects attention for edited queries, suppressing object synthesis. Our Adaptive Context Enrichment resolves this by offering a dual context: the Editing Path for synthesizing the new object, and the Reconstruction Path for preserving background structure. Attention in unedited regions remains correct, confirming our method is non-invasive.

Adaptive Fusion via Context Enrichment Our solution is fundamentally different: adaptive context enrichment. Instead of replacing the context, we enrich it, empowering the pre-trained attention module to perform a dynamic, content-aware fusion. To implement this, we access both editing and reconstruction contexts simultaneously via a synchronized dual-path process. Both paths originate from the same noise anchor $\mathbf{z_1}$. The first, our Reconstruction Path, is conditioned on the original video inputs $(V_{src}[0], s_\phi)$. It focuses on preserving content fidelity by denoising $\mathbf{z_1}$ back to the source video, providing the essential source context (keys $K_{t,l}^{res}$ and values $V_{t,l}^{res}$). In parallel, the Editing Path handles the creative task. Conditioned on the edit inputs (I_{edit}, s_{tgt}), it synthesizes the desired changes from the same $\mathbf{z_1}$, providing the editing queries $Q_{t,l}^{edit}$ and its own internal context $(K_{t,l}^{edit}, V_{t,l}^{edit})$.

With both sets of contexts available, we perform the enrichment within the Editing Path's self-attention. We augment the key and value by concatenating them with their counterparts from the Reconstruction Path:

$$K_{aug} = \text{Concat}([K_{t,l}^{edit}, K_{t,l}^{res}]) \tag{4}$$

$$V_{aug} = \text{Concat}([V_{t,l}^{edit}, V_{t,l}^{res}]) \tag{5}$$

The self-attention is then computed using them:

$$\text{Self-Attn}' = \text{softmax}\left(\left[Q_{t,l}^{edit}(K_{aug})^T\right]/\sqrt{d}\right) V_{aug} \tag{6}$$

This design leverages the inherent optimization behavior of self-attention. The enriched key-value space allows each query to attend to its most relevant information—be it from the source context for background preservation, or the edit context for new content synthesis. This transforms guidance from a rigid command into a dynamically weighted fusion process, enabling a robust and high-fidelity fusion.

Targeted Fusion via Vital Layer Analysis

Having established *how* to guide the generation, we now address the critical question of *where*. Injecting guidance uniformly across all layers of a DiT is not only computationally wasteful but also conceptually flawed. Applying

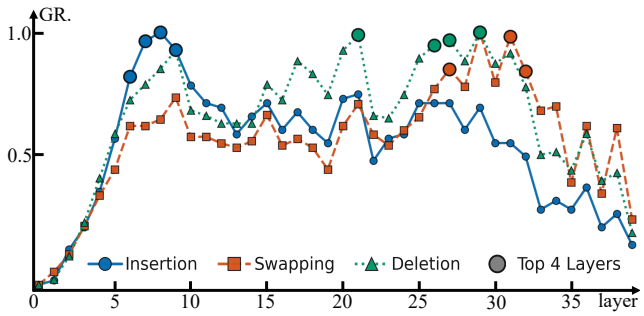

Figure 5: Task-Dependent Guidance Responsiveness (min-max normalized data in the figure). A higher Guidance Responsiveness indicates greater influence, and there are three primary zones across all layers.

our semantic-level guidance uniformly across all layers risks disrupting the DiT's functional hierarchy of layers, potentially weakening the edit's results. Therefore, a targeted intervention is required. Prior works on U-Net architectures have relied on empirical heuristics for layer selection. However, such heuristics are not reliably transferable to the different and more homogeneous structure of DiTs.

To formalize this, we propose a data-driven method to identify the most influential layers. We define a Guidance Responsiveness Metric, GR_l, that quantifies a layer's responsiveness to our Contextual Enrichment mechanism. For each layer l, we calculate this by performing a one-step denoising on a set of videos with the editing conditions and computing two feature maps for each layer: $x_l^{\text{no-CE}}$ (i.e., layer l's self-attention output without Contextual Enrichment) and x_l^{CE}(i.e., layer l's self-attention output with Contextual Enrichment applied only at layer l). The Guidance Responsiveness is measured by the dissimilarity:

$$GR_l = 1 - \text{mean}\left(\text{cosine\_similarity}\left(x_l^{\text{no-CE}}, x_l^{\text{CE}}\right)\right) \quad (7)$$

A high GR_l score signifies that the layer is highly sensitive to the guidance and thus influential in the editing process. Applying this analysis across a range of editing tasks reveals that layer responsiveness is not uniform, but instead exhibits a highly structured and task-dependent pattern, as shown in Figure 5. There are three primary zones of high responsiveness across the model's depth: an early block (layers 1-10), a middle block (layers 15-21), and a deep block (layers 26-32). Crucially, the dominant zone of activity varies systematically with the task. For object insertion, peak responsiveness is consistently located in the early-layer block. In contrast, object swapping elicits the strongest response in the deep-layer block. Object deletion presents a unique dual-peak pattern, showing high responsiveness in both the middle and deep blocks. These empirical evidence strongly suggests a structural pattern within the DiT's layers. The observed patterns are consistent with the broader understanding of Transformer architectures, where early layers typically handle spatial and structural information, while deeper layers manage more abstract semantic concepts (Tenney, Das, and Pavlick 2019; Raghu et al. 2021). For instance, the reliance of insertion on early layers aligns with a need to establish spatial layout, while deletion's dependence on deep layers corresponds to a high-level semantic operation.

By selecting only the top-k layers with the highest importance for each task, we ensure our intervention is potent, targeted, and computationally efficient. This principled selection strategy completes our framework, delivering a robust and precise video editing solution.

4 Experiments

Experimental Setup

Implementation Details Our framework is built upon Wan2.1-I2V-14B-480P (Wan et al. 2025), a publicly available image-to-video Diffusion Transformer with 40 layers. We adhere to a training-free paradigm, requiring no optimization or fine-tuning of the pre-trained diffusion model. For inversion, we utilize RF-Solver (Wang et al. 2025b) with 50 steps to map the source video into a noise latent, and the subsequent editing also uses 50 sampling steps. We set the timestep threshold τ to 0.5 (i.e., the mechanism is adopted for the first 50% of timesteps) and a guidance scale of 3.0.

Evaluation Dataset and Baselines We evaluate our method on the Unic-Benchmark (Ye et al. 2025). For first-frame editing, we use AnyDoor (Chen et al. 2024) for insertion, InsertAnything (Song et al. 2025) for swapping, and MagicQuill (Liu et al. 2025b) for deletion. We compare against baselines including training-free AnyV2V (Ku et al. 2024), our adapted AnyV2V-DiT, and training-based VACE (Jiang et al. 2025) and I2VEdit (Ouyang et al. 2024). For evaluation, we measure task-specific fidelity (CLIP-score, DINO-score), background preservation (PSNR, SSIM), and overall video quality using metrics from VBench (Huang et al. 2024) (e.g., consistency, smoothness, dynamic and aesthetic quality).

Comparison with Baselines

We conduct a comprehensive quantitative evaluation against state-of-the-art methods on object insertion, swapping, and deletion tasks. The results, summarized in Table 1 and Figure 6, demonstrate the performance of ContextFlow.

For creative edits like object insertion and swapping, ContextFlow outperforms most baselines in critical Identity and Alignment metrics. Its leading Aesthetic and Dynamic scores also reflect high visual fidelity and temporal coherence. In contrast, while I2VEdit achieves higher Identity scores, it requires per-video training and produces stiff, copy-paste-like results. Other methods show more significant flaws: VACE fails on out-of-distribution insertion and exhibits poor identity matching in swaps, while AnyV2V yields blurry visuals and unstable objects.

In the deletion task, ContextFlow achieves top scores in Smoothness and Dynamics while maintaining high reconstruction quality. In contrast, VACE incorrectly replaces objects with hallucinated content, while AnyV2V leaves shadow artifacts and shows poor reconstruction quality.

The "AnyV2V-DiT" baseline shows that directly applying traditional U-Net guidance mechanisms to the DiT architecture causes severe visual artifacts, including geometric

Figure 6: Qualitative comparisons of our method against open-source video editing baselines. The guiding text prompt is shown below the original videos. Our method demonstrates satisfactory results. Zoom in for better visualization.

Task	Method	Identity		Alignment		Video Quality			Reconstruction Quality	
		$CLIP_I$↑	$DINO_I$↑	$CLIP_{score}$↑	Overall cons.↑	Smooth↑	Dynamic↑	Aesthetic↑	PSNR↑	SSIM↑
Insert	AnyV2V	0.5943	0.4053	0.2776	0.1887	0.9804	0.3077	0.5287	20.57	0.7055
	VACE	0.5683	0.3967	0.2569	0.1386	**0.9921**	0.3077	0.5724	18.86	**0.9033**
	AnyV2V-DiT	0.6376	0.4479	0.3060	0.2579	0.9917	0.3846	0.6145	26.06	0.8478
	I2VEdit	**0.6710**	**0.4595**	**0.3124**	0.2600	0.9827	0.3077	0.5846	26.23	0.8360
	Ours	0.6504	0.4566	0.3107	**0.2691**	0.9918	**0.4231**	**0.6227**	**26.26**	0.8575
Swap	AnyV2V	0.6046	0.5641	0.3210	0.2384	0.9848	0.0769	0.5739	21.88	0.6867
	VACE	0.6080	0.5917	0.3226	0.2412	**0.9926**	**0.1538**	0.6144	**29.63**	**0.9238**
	AnyV2V-DiT	0.6617	0.5983	0.3362	0.2597	0.9907	**0.1538**	0.6076	20.18	0.6854
	I2VEdit	**0.6683**	0.6003	0.3282	0.2595	0.9819	0.0769	0.5995	26.19	0.7966
	Ours	0.6644	**0.6004**	**0.3391**	**0.2648**	0.9924	0.0769	**0.6176**	22.66	0.7518
Delete	AnyV2V	–	–	**0.2891**	0.2170	0.9781	0.1500	0.5378	22.07	0.6530
	VACE	–	–	0.2794	0.1948	0.9889	0.3000	**0.5645**	**31.57**	**0.9064**
	AnyV2V-DiT	–	–	0.2863	**0.2136**	0.9886	0.3000	0.5413	21.14	0.6514
	I2VEdit	–	–	0.2790	0.2081	0.9816	0.3000	0.5169	25.44	0.7758
	Ours	–	–	0.2854	0.2111	**0.9900**	**0.3500**	0.5405	22.16	0.7030

Table 1: Quantitative comparison on object insertion, swap, and deletion tasks. Our method has achieved impressive performance across numerous metrics for each task, demonstrating its comprehensiveness.

distortions, inconsistent temporal dynamics, and spatial deformations. This result provides empirical evidence that the architectural transition to Transformers requires fundamentally new guidance mechanisms.

Ablation Studies

To rigorously analyze the contributions of our proposed components, we conduct a series of ablation studies on the object insertion task. Our analysis is structured to answer three fundamental questions regarding our ContextFlow framework: 1) Is our guidance mechanism effective and how should it be implemented? 2) How much guidance is optimal and where should it be injected? 3) During which phase of the denoising process should guidance be active?

Core Mechanism Validation: Guidance Strategy We first validate our adaptive Context Enrichment (CE) mechanism. We compare our method against two critical variants: one that omits CE entirely, relying solely on the inverted noise and edited frame, and another that substitutes our K/V concatenation with a conventional replacement strategy.

As presented in Table 2, ablating the CE module leads to a discernible decline across all metrics. While high-fidelity inversion provides a strong foundation, explicit guidance during the denoising path is crucial for context-aware editing. More revealingly, the "K/V Replacement" strategy significantly impairs Identity Preservation scores. We attribute this to the destructive nature of replacement, which discards valuable contextual information from the editing path's context. In contrast, our concatenation approach is additive; it enriches the context, empowering the self-attention module to dynamically balance between source and target contexts.

Method	$\text{CLIP}_\text{I}\uparrow$	$\text{DINO}_\text{I}\uparrow$	$\text{CLIP}_\text{score}\uparrow$	Overall.↑	Aesth.↑
Ours	**0.6504**	**0.4566**	**0.3107**	**0.2691**	**0.6227**
w/o CE	0.6447	0.4529	0.3086	0.2634	0.6186
KV Replace	0.6349	0.4508	0.3018	0.2544	0.6200

Table 2: Ablation on the core guidance strategy. Our method demonstrates clear superiority over both the absence of guidance and a destructive replacement approach.

Targeted Guidance Analysis: How Much and Where? We now investigate the specifics of our method: determining the optimal *quantity* and *location* for K/V injection.

How Much Guidance? We first analyze the impact of $\mathbf{k}$, the number of top-ranked layers selected for injection. As detailed in Table 3, the results reveal a distinct unimodal performance curve. Insufficient guidance ($k < 4$) fails to provide a robust structural anchor, leading to weaker identity preservation. Conversely, excessive guidance ($k > 4$) over-constrains the model, stifling its generative capacity and causing the desired edit to diminish, as evidenced by the sharp performance degradation for $k = 32$ and $k = 40$. An optimal balance is achieved at $k = \mathbf{4}$, which corresponds to the top 10% of layers in the 40-layer DiT.

k	$\text{CLIP}_\text{I}\uparrow$	$\text{DINO}_\text{I}\uparrow$	$\text{CLIP}_\text{score}\uparrow$	Overall.↑	Aesth.↑
0	0.6447	0.4529	0.3086	0.2634	0.6186
1	0.6467	0.4537	0.3087	0.2645	0.6196
2	0.6452	0.4530	0.3077	0.2623	0.6188
4	**0.6504**	**0.4566**	**0.3107**	**0.2691**	0.6227
8	0.6456	0.4541	0.3089	0.2620	**0.6240**
16	0.6330	0.4435	0.3058	0.2566	0.6157
32	0.6100	0.4303	0.2959	0.2458	0.5998
40	0.5715	0.4067	0.2685	0.1863	0.5863

Table 3: Ablation on the number of injected layers (k). Performance peaks at $k = 4$, which is an optimal trade-off between guidance strength and generative freedom.

Where to Inject? We now validate our principled method for selecting *which* four layers to target. In Table 4, we benchmark our Guidance Responsiveness (GR)-based selection against several alternatives. Injecting into all 40 layers proves detrimental, corroborating our earlier finding that over-constraining the model is counterproductive. Conversely, injecting into the four *least*-responsive layers provides insufficient structural cues, failing to anchor the edit effectively. Finally, a U-Net-based heuristic adapted from AnyV2V (layers 3-10) performs commendably but is still surpassed by our method on key identity metrics. This demonstrates that our data-driven Guidance Responsiveness metric is not merely a theoretical construct but a practical and superior tool for identifying the most impactful layers for precise video editing.

Method	$\text{CLIP}_\text{I}\uparrow$	$\text{DINO}_\text{I}\uparrow$	$\text{CLIP}_\text{score}\uparrow$	Overall.↑	Aesth.↑
Ours	**0.6504**	**0.4566**	0.3107	**0.2691**	**0.6227**
Injection on:					
①	0.5715	0.4067	0.2685	0.1863	0.5863
②	0.6138	0.4367	0.2970	0.2365	0.6059
③	0.6458	0.4553	**0.3109**	0.2634	0.6210

Table 4: Ablation on the layer selection strategy. ①: all layers; ②: 4 least-responsive layers; ③: layers follow AnyV2V. Our principled selection outperforms other approaches, underscoring the effectiveness of our GR metric.

Guidance Window Analysis: When to Inject? Finally, we analyze the injection timestep threshold τ, which governs the temporal duration of the Contextual Enrichment mechanism. This parameter critically mediates the trade-off between structural preservation (favoring higher τ) and edit flexibility (favoring lower τ). As evaluated in Table 5, a low $\tau = 0.2$ offers insufficient guidance, while a high $\tau = 1.0$, which applies guidance throughout the entire process, slightly compromises aesthetic quality by restricting the model during the final, high-fidelity refinement stages. We identify $\tau = 0.5$ as the optimal equilibrium.

τ	$\text{CLIP}_\text{I}\uparrow$	$\text{DINO}_\text{I}\uparrow$	$\text{CLIP}_\text{score}\uparrow$	Overall.↑	Aesth.↑
0.2	0.6465	0.4547	0.3098	0.2600	0.6200
0.5	**0.6504**	0.4566	0.3107	0.2691	**0.6227**
1.0	0.6482	**0.4568**	**0.3115**	**0.2713**	0.6154

Table 5: Ablation on the injection timestep τ. A value of $\tau = 0.5$ strikes the best balance between edit fidelity and structural coherence.

5 Conclusion

We present ContextFlow, a training-free framework for video object editing. Our core contribution, Adaptive Context Enrichment, injects controlled structural guidance from a parallel reconstruction path via a Key-Value concatenation mechanism. By targeting vital layers identified through Guidance Responsiveness analysis, our method balances edit fidelity with background stability. Experiments confirm the effectiveness of our approach in creating high-quality, consistent video object edits, empowering users with fine-grained creative control.

References

Bai, J.; He, T.; Wang, Y.; Guo, J.; Hu, H.; Liu, Z.; and Bian, J. 2024. UniEdit: A Unified Tuning-Free Framework for Video Motion and Appearance Editing. *Proceedings of the 33rd ACM International Conference on Multimedia*.

Burgert, R.; Xu, Y.; Xian, W.; Pilarski, O.; Clausen, P.; He, M.; Ma, L.; Deng, Y.; Li, L.; Mousavi, M.; Ryoo, M.; Debevec, P.; and Yu, N. 2025. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13–23.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2024. AnyDoor: Zero-shot Object-level Image Customization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6593–6602.

Gao, C.; Ding, L.; Cai, X.; Huang, Z.; Wang, Z.; and Xue, T. 2025. LoRA-Edit: Controllable First-Frame-Guided Video Editing via Mask-Aware LoRA Fine-Tuning. arXiv:2506.10082.

Geyer, M.; Bar Tal, O.; Bagon, S.; and Dekel, T. 2024. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *International Conference on Representation Learning*, volume 2024, 1608–1620.

Graikos, A.; Malkin, N.; Jojic, N.; and Samaras, D. 2022. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35: 14715–14728.

Gu, Y.; Zhou, Y.; Wu, B.; Yu, L.; Liu, J.-W.; Zhao, R.; Wu, J. Z.; Zhang, D. J.; Shou, M. Z.; and Tang, K. 2024. VideoSwap: Customized Video Subject Swapping with Interactive Semantic Point Correspondence. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7621–7630.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; Wang, Y.; Chen, X.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21807–21818.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2024. VMC: Video Motion Customization Using Temporal Attention Adaption for Text-to-Video Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9212–9221.

Jiang, Z.; Han, Z.; Mao, C.; Zhang, J.; Pan, Y.; and Liu, Y. 2025. VACE: All-in-One Video Creation and Editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17191–17202.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6507–6516.

Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *Transactions on Machine Learning Research*.

Liu, S.; Wang, T.; Wang, J.-H.; Liu, Q.; Zhang, Z.; Lee, J.-Y.; Li, Y.; Yu, B.; Lin, Z.; Kim, S. Y.; and Jia, J. 2025a. Generative Video Propagation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17712–17722.

Liu, Z.; Yu, Y.; Ouyang, H.; Wang, Q.; Cheng, K. L.; Wang, W.; Liu, Z.; Chen, Q.; and Shen, Y. 2025b. MagicQuill: An Intelligent Interactive Image Editing System. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13072–13082.

Molad, E.; Horwitz, E.; Valevski, D.; Acha, A. R.; Matias, Y.; Pritch, Y.; Leviathan, Y.; and Hoshen, Y. 2023. Dreamix: Video Diffusion Models are General Video Editors. arXiv:2302.01329.

Mou, C.; Cao, M.; Wang, X.; Zhang, Z.; Shan, Y.; and Zhang, J. 2024. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37: 18481–18505.

Ouyang, W.; Dong, Y.; Yang, L.; Si, J.; and Pan, X. 2024. I2VEdit: First-Frame-Guided Video Editing via Image-to-Video Diffusion Models. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15886–15896.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do Vision Transformers See Like Convolutional Neural Networks? In *Advances in Neural Information Processing Systems*, volume 34, 12116–12128.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22500–22510.

Saini, N.; Bodla, N.; Shrivastava, A.; Ravichandran, A.; Zhang, X.; Shrivastava, A.; and Singh, B. 2024. InVi: Object Insertion In Videos Using Off-the-Shelf Diffusion Models. arXiv:2407.10958.

Shen, Z.; Wu, C.; Zhou, J.; Zhao, C.; Wang, K.; Zhou, H.; Li, Y.; Feng, H.; He, W.; and Wang, J. 2025. iDiT-HOI: Inpainting-based Hand Object Interaction Reenactment via Video Diffusion Transformer. arXiv:2506.12847.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*, volume 2021.

Song, W.; Jiang, H.; Yang, Z.; Quan, R.; and Yang, Y. 2025. Insert Anything: Image Insertion via In-Context Editing in DiT. arXiv:2504.15009.

Su, T.; Wang, C.; Huang, J.; and Lu, D. 2025. Zero-to-Hero: Zero-Shot Initialization Empowering Reference-Based Video Appearance Editing. arXiv:2505.23134.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

Tu, Y.; Luo, H.; Chen, X.; Ji, S.; Bai, X.; and Zhao, H. 2025. VideoAnydoor: High-fidelity Video Object Insertion with Precise Motion Control. arXiv:2501.01427.

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.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314.

Wang, G.; Fan, S.; Liu, H.; Song, Q.; Wang, H.; and Xu, J. 2025a. Consistent Video Editing as Flow-Driven Image-to-Video Generation. arXiv:2506.07713.

Wang, J.; Pu, J.; Qi, Z.; Guo, J.; Ma, Y.; Huang, N.; Chen, Y.; Li, X.; and Shan, Y. 2025b. Taming Rectified Flow for Inversion and Editing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 64044–64058.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024. MotionCtrl: A Unified and Flexible Motion Controller for Video Generation. In *ACM SIGGRAPH 2024 Conference Papers*, SIGGRAPH 2024.

Wei, Y.; Zhang, S.; Yuan, H.; Wang, X.; Qiu, H.; Zhao, R.; Feng, Y.; Liu, F.; Huang, Z.; Ye, J.; Zhang, Y.; and Shan, H. 2024. DreamVideo-2: Zero-Shot Subject-Driven Video Customization with Precise Motion Control. arXiv:2410.13830.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7589–7599.

Wu, T.; Zhang, Y.; Wang, X.; Zhou, X.; Zheng, G.; Qi, Z.; Shan, Y.; and Li, X. 2025. CustomCrafter: Customized Video Generation with Preserving Motion and Concept Composition Abilities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8): 8469–8477.

Ye, Z.; He, X.; Liu, Q.; Wang, Q.; Wang, X.; Wan, P.; Zhang, D.; Gai, K.; Chen, Q.; and Luo, W. 2025. UNIC: Unified In-Context Video Editing. arXiv:2506.04216.

Zhuang, S.; Huang, Z.; Yang, B.; Zhang, Y.; Wang, F.; Fu, C.; Sun, C.; Zha, Z.-J.; Li, C.; and Wang, Y. 2025. Get In Video: Add Anything You Want to the Video. arXiv:2503.06268.