# Continuous Control of Editing Models via Adaptive-Origin Guidance

ALON WOLF, Tel Aviv University, Decart.ai, Israel
CHEN KATZIR, Decart.ai, Israel
KFIR ABERMAN, Decart.ai, Israel
OR PATASHNIK, Tel Aviv University, Israel

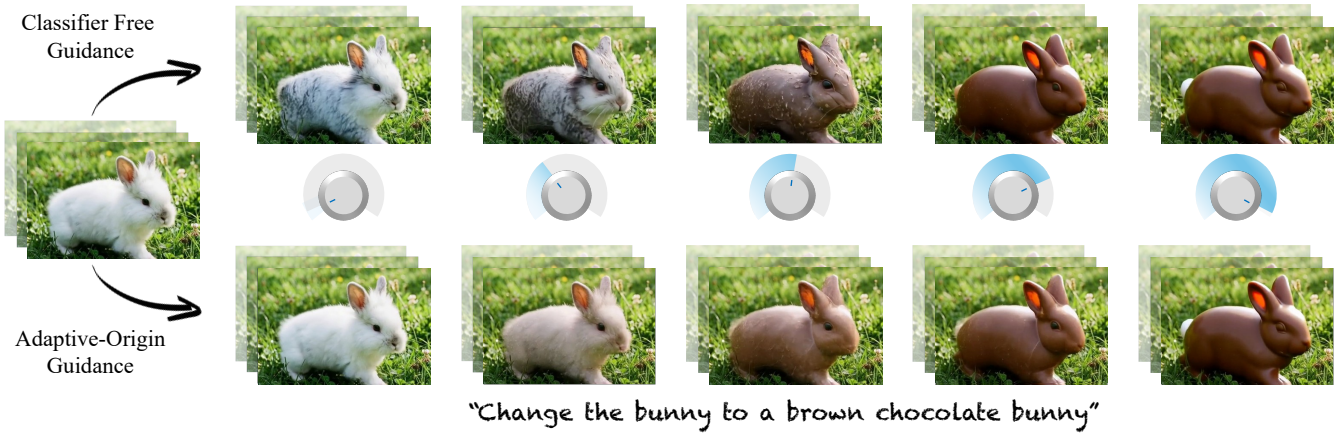



Fig. 1. Given an input video (left) and a text instruction ("change the bunny to a brown chocolate bunny"), standard Classifier-Free Guidance (top row) does not recover the input content as the guidance scale decreases. Instead of a smooth transition back to the source, it yields arbitrary, inconsistent modifications (e.g., the gray bunny on the left). In contrast, our Adaptive-Origin Guidance (bottom row) enables a controllable gradual, semantically consistent progression editing from the original video to the fully edited result by adaptively interpolating the guidance origin.

Diffusion-based editing models have emerged as a powerful tool for semantic image and video manipulation. However, existing models lack a mechanism for smoothly controlling the intensity of text-guided edits. In standard text-conditioned generation, Classifier-Free Guidance (CFG) impacts prompt adherence, suggesting it as a potential control for edit intensity in editing models. However, we show that scaling CFG in these models does not produce a smooth transition between the input and the edited result. We attribute this behavior to the unconditional prediction, which serves as the guidance origin and dominates the generation at low guidance scales, while representing an arbitrary manipulation of the input content. To enable continuous control, we introduce Adaptive-Origin Guidance (AdaOr), a method that adjusts this standard guidance origin with an identity-conditioned adaptive origin, using an identity instruction corresponding to the identity manipulation. By interpolating this identity prediction with the standard unconditional prediction according to the edit strength, we ensure a continuous transition from the input to the edited result. We evaluate our method on image and video editing tasks, demonstrating that it provides smoother and more consistent control compared to current slider-based editing approaches. Our method incorporates an identity instruction into the standard training framework, enabling fine-grained control at inference time without per-edit procedure or reliance on specialized datasets. Additional results and videos are available at https://adaor-paper.github.io/.

## 1 Introduction

Recent years have witnessed significant progress in diffusion models for the creation and manipulation of visual content [Ho et al. 2020; Rombach et al. 2022; Song et al. 2022; Wan et al. 2025]. In particular, in the realm of image and video editing, these models enable complex semantic manipulations guided by natural language [DecartAI 2025; Hertz et al. 2022; Labs et al. 2025]. However, while text prompts offer an intuitive interface for specifying what should be changed, they provide limited control over how much the change should be applied. For instance, when editing a portrait to add a beard, one may wish to smoothly vary the result from a clean-shaven appearance to light stubble and eventually to a full beard, rather than committing to a single, fixed outcome. Supporting such gradual transitions between the source and the edited result, however, remains challenging for existing diffusion-based editing models.

While recent research has explored methods for achieving edit strength control, these approaches often rely on per-edit-type procedures [Gandikota et al. 2023, 2025] or require extensive data collection [Parihar et al. 2025a]. As a result, they tend to lack robustness across different inputs, edits, and model architectures, and their applicability to video editing remains largely unexplored.

In text-conditioned generation, Classifier-Free Guidance (CFG) is a crucial mechanism for ensuring visual quality and alignment between the prompt and the generated content [Ho and Salimans 2022]. By scaling the guidance, the model is forced to adhere more strictly to the conditioning signal, effectively increasing the influence of the prompt on the final output. Given the central role of CFG in modulating prompt influence in both text-to-image and text-to-video generation, a natural question is whether adjusting the CFG scale can provide smooth control over edit strength in diffusion-based editing models. We show that this is not the case: in editing models, lowering the CFG scale does not produce a smooth transition back to the unedited input (see Figures 1 and 2).

The key observation of this work is that the limitation of CFG in controlling editing strength arises from the dominance of the unconditional prediction at low guidance scales and from the nature

of this prediction in editing models. We refer to this term as the *guidance origin*. In instruction-based editing settings, the unconditional prediction typically corresponds to an arbitrary manipulation of the input rather than faithful reconstruction. Consequently, when the guidance scale is varied, low guidance values do not induce small semantic changes around the input. Instead, the denoising process becomes dominated by the unconditional prediction serving as the guidance origin, leading to arbitrary deviations from the source content and steering the edit along uncontrolled directions.

To enable smooth control over edit strength in diffusion-based editing models, we introduce *Adaptive Origin Guidance* . We first propose an identity instruction, an instruction that corresponds to the identity manipulation: reproducing the input content without any semantic modification. Building on this, we introduce a guidance mechanism where the term that dominates the prediction at low scales (i.e., the origin) is adjusted according to the desired edit strength. Specifically, we interpolate between the identity prediction and the standard unconditional prediction. By assigning greater weight to the identity term at lower edit strengths and transitioning to the standard term at higher strengths, our method enables smooth, continuous control over manipulation intensity.

We evaluate our method on both image and video editing models, demonstrating its effectiveness across various architectures and manipulation tasks. Our results show that the proposed guidance strategy provides smooth, high-quality transitions that are significantly more consistent than standard CFG-based scaling. Furthermore, we compare our approach against existing slider-based editing methods [Gandikota et al. 2023; Kamenetsky et al. 2025; Parihar et al. 2025a], demonstrating superior performance in balancing semantic change with structural preservation. Our guidance mechanism enables precise control over manipulation intensity without requiring per-edit optimization or the collection of specialized datasets.

## 2 Related Work

***Instruction-driven Image and Video Editing.*** In recent years, the ability to generate and edit visual content using natural language has advanced rapidly [Nichol et al. 2022; Patashnik et al. 2021; Ramesh et al. 2022; Saharia et al. 2022]. In particular, progress in text-conditioned diffusion models has enabled high-quality semantic manipulations that previously required substantial manual effort [Avrahami et al. 2022; Brooks et al. 2023; Hertz et al. 2022; Meng et al. 2022; Tumanyan et al. 2023]. Early works primarily focused on image editing, with a prominent line of research steering the denoising process in a training-free manner using feature injection [Alaluf et al. 2023; Avrahami et al. 2025; Cao et al. 2023; Garibi et al. 2024; Hertz et al. 2022; Mokady et al. 2022; Parihar et al. 2025b; Patashnik et al. 2023; Tumanyan et al. 2023], latent optimization [Epstein et al. 2023; Parmar et al. 2023], or alternative sampling strategies [Huberman-Spiegelglas et al. 2023; Kulikov et al. 2025; Meng et al. 2022; Rout et al. 2025]. Several of these approaches have later been generalized to video editing [Gao et al. 2025; Geyer et al. 2023; Ku et al. 2024; Wu et al. 2023; Yatim et al. 2025].

More recently, instruction-driven diffusion models have demonstrated state-of-the-art performance in image editing [Labs 2025; Labs et al. 2025; Team 2025; Wu et al. 2025b]. These models take as input an image and a text instruction, and produce a corresponding edited image [Brooks et al. 2023]. A central challenge in training such models lies in data collection, as paired examples of an image and its edited version are rarely available in the wild. As a result, training-free editing methods are often used to synthetically construct datasets for supervised training. With the rapid advancement of video generation models [HaCohen et al. 2026, 2024; Kong et al. 2024; Wan et al. 2025], several video editing models have also been introduced [DecartAI 2025; Jiang et al. 2025], including Lucy [DecartAI 2025], an instruction-driven video editing model.

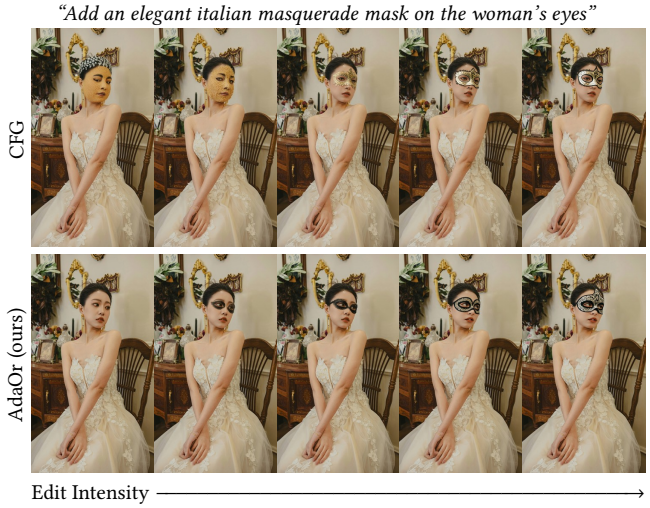



Fig. 2. **Comparison with standard CFG scaling.** We compare the progression of the edit as the intensity increases from 0 (left) to 1 (right). Standard CFG (top) exhibits arbitrary modifications at lower scales (e.g., the gold face paint). In contrast, our method (bottom) enables a smooth, continuous transition, gradually introducing the mask features while maintaining the structural integrity of the face throughout the interpolation.

***Continuous Control for Image Editing.*** While natural language provides a form of semantic control that is difficult to achieve using other modalities, it is inherently coarse. As a result, achieving fine-grained control over generated or edited content using language alone is often challenging. In particular, continuous control over edit strength is difficult to express through text. To address this limitation, recent works introduce slider-based controls into generative models, enabling explicit manipulation of edit strength via a continuous parameter [Dravid et al. 2024; Gandikota et al. 2023, 2025; Li et al. 2024; Parihar et al. 2025a, 2024].

Among these approaches, some rely on per-edit-type procedures, such as optimization or direction construction [Gandikota et al. 2023, 2025; Kamenetsky et al. 2025], which incur additional computational cost for each edit type. Other methods operate in the text embedding space, identifying editing directions and traversing them with varying step sizes corresponding to different edit strengths [Baumann et al. 2025; Dalva et al. 2024; Kamenetsky et al. 2025]. Another line of work collects datasets containing edits at multiple strengths and trains models that are explicitly conditioned on the edit strength [Cheng et al. 2025; Magar et al. 2025; Parihar et al. 2025a]. Since collecting such datasets is challenging, most of these methods focus on a narrow class of edits, such as material changes. Some works [Parihar et al. 2025a; Xu et al. 2025] construct datasets

that are not restricted to a specific domain, but require substantial computational resources for data generation and model training. In contrast, our method does not rely on a custom dataset. As a result, it supports the diverse range of edit types already handled by the backbone editing model. Notably, existing approaches focus exclusively on image editing and do not address video.

Another related line of work is image morphing [Cao et al. 2025; Zhang et al. 2024]. These methods take two images as input and generate a continuous sequence that morphs between them. When combined with image editing, they can interpolate between the original and edited images. However, they often exhibit noticeable jumps, as they operate in latent spaces that lack sufficient continuity, and they remain largely underexplored in the video domain.

***Guidance in Diffusion Models.*** To achieve high-fidelity alignment with the conditioning prompt, text-conditioned diffusion models rely on inference-time mechanisms to steer the generative process. Early methods relied on external classifiers to guide the generation process using the gradient of the classifier with respect to the input to modify the diffusion model's denoising prediction [Dhariwal and Nichol 2021]. While effective, this approach introduces additional complexity by requiring external models at inference.

Classifier-Free Guidance (CFG) [Ho and Salimans 2022] removes the need for a separate classifier by using the diffusion model itself as an "implicit classifier". This is achieved by modifying the training procedure so that the model is trained not only as a conditional generator, but also as an unconditional one. Specifically, unconditional behavior is learned by training the model to denoise arbitrary inputs when the null condition (typically the empty string, denoted as $\varnothing$) is provided. Having access to both conditional and unconditional predictions allows these outputs to be combined at inference time in a way that is equivalent to using a classifier, enabling the diffusion model to act as this implicit classifier.

While the original derivation of CFG was grounded in the gradient of the log-likelihood of an implicit classifier, follow-up works have increasingly analyzed this mechanism through different lenses, proposing various interpretations and improvements [Bradley and Nakkiran 2024; Chung et al. 2025; Hyung et al. 2025; Karras et al. 2024; Katzir et al. 2024; Yehezkel et al. 2025].

Recently, instruction-based editing models have become increasingly popular for both image and video editing [DecartAI 2025; Labs et al. 2025; Wu et al. 2025b]. These models rely on two conditioning signals: the input image or video to be edited and the edit instruction. Similarly to text-conditioned models, these models also employ CFG to enforce high prompt adherence. While earlier approaches [Brooks et al. 2023] utilized dual guidance by independently dropping both the text and the input image, recent state-of-the-art editing models typically restrict guidance to the instruction alone.

## 3 Method

### 3.1 Preliminaries

Classifier-free guidance (CFG) combines conditional and unconditional noise predictions to steer the generation process. Given a noisy latent variable $\mathbf{z}_t$ at diffusion timestep $t$, the guided noise prediction is defined as:

$$\epsilon^w(\mathbf{z}_t; c, t) \;=\; \epsilon(\mathbf{z}_t; \varnothing, t) \;+\; w\big(\epsilon(\mathbf{z}_t; c, t) - \epsilon(\mathbf{z}_t; \varnothing, t)\big),$$

where $\epsilon(\cdot)$ denotes the noise prediction network of the diffusion model, $c$ is the conditioning signal (e.g., a text prompt), and $\varnothing$ denotes the unconditional (null) input. The scalar $w \geq 0$ is the guidance scale controlling the strength of conditioning, with larger values encouraging closer adherence to $c$.

Geometrically, noise predictions can be viewed as vectors inducing transitions between successive noise distributions, namely, transitions from the marginal noise distribution $p_t$ at time $t$ to the slightly less noisy distribution $p_{t-1}$. Under this view, CFG decomposes into an *origin* given by the unconditional score, $\epsilon(\mathbf{z}_t; \varnothing, t)$, which moves the latent from $p_t$ toward the manifold of $p_{t-1}$, and a *steering component* $\epsilon(\mathbf{z}_t; c, t) - \epsilon(\mathbf{z}_t; \varnothing, t)$ that acts on this manifold to bias the trajectory toward the conditional distribution, as illustrated in Figure 3a.

For *instruction-based editing models*, which are conditioned on an input image or video $c_I$ and an edit instruction $c_T$, CFG is commonly used to enforce prompt adherence. In practice, guidance is typically applied only to the instruction, and models are trained to handle a null instruction $\varnothing$ by randomly dropping $c_T$ during training while still denoising the target output conditioned on $c_I$. At inference, the guided prediction at timestep $t$ is:

$$\epsilon^w(\mathbf{z_t}; c_I, c_T, t) = \epsilon(\mathbf{z_t}; c_I, \varnothing, t) + w(\epsilon(\mathbf{z_t}; c_I, c_T, t) - \epsilon(\mathbf{z_t}; c_I, \varnothing, t)),$$

with $w$ controlling the strength of instruction guidance.

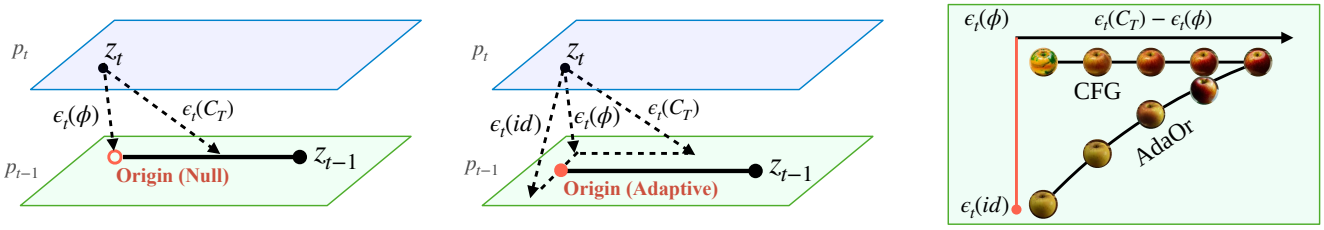
### 3.2 Adaptive-Origin Guidance

Motivated by the fact that the guidance scale $w$ in CFG controls prompt adherence, our goal is to leverage it to control edit strength. However, as demonstrated in Figures 1 and 2, this approach does not work naïvely. We first observe that for small values of $w$, the noise prediction $\epsilon^w(\mathbf{z}_t; c_I, c_T, t)$ is dominated by the origin term $\epsilon(\mathbf{z}_t; c_I, \varnothing, t)$, rather than by the steering direction $\epsilon(\mathbf{z}_t; c_I, c_T, t) - \epsilon(\mathbf{z}_t; c_I, \varnothing, t)$. In other words, as $w \to 0$, the prediction collapses to the guidance origin.

This raises the question: what is the semantic meaning of the null prediction in an editing model? We observe that the null instruction $\varnothing$ effectively functions as an "any edit" command. This parallels the role of the null prompt in standard text-conditioned models, which represents the distribution of "any natural image". In the context of editing, however, since the model is conditioned on an input image, the null term corresponds to the marginal distribution of valid edits. As a result, it projects the input onto a generic manifold of edited images. We demonstrate such generic edits at small CFG scales in Figure 1, where the bunny turns gray instead of brown, and in Figure 2, where the face is painted gold and a hair decoration is added.

This observation explains why standard CFG fails to provide continuous control over edit strength. Intuitively, lowering the guidance scale ($w \to 0$) should reduce the edit magnitude and generate the input image. However, because the standard origin represents an "any edit" point, it drives the output toward a generic manifold of edited images (see Figure 3a). As a result, small values of $w$ produce arbitrary edits rather than weak ones, causing the output to lose input-specific structure in favor of generic dataset features.

To achieve continuous control over edit strength, we propose an adaptive-origin guidance that interpolates between a "no-edit"




(a) **Standard CFG**: Null-condition as origin

(b) **Adaptive Origin**: Null-identity interpolated origin

(c) **Edit Progression Comparison**: Varying CFG scale vs AdaOr edit strength

Fig. 3. **Geometric interpretation of standard Classifier-Free Guidance and Adaptive Origin Guidance.** In both (a) and (b), we illustrate a single denoising step in which the latent $\mathbf{z}_t$ lies on the manifold of the marginal distribution $p_t$. The origin prediction first denoises $\mathbf{z}_t$ toward the manifold of the less noisy distribution $p_{t-1}$, after which the trajectory is steered on this manifold toward better alignment with the conditioning signal. In (a) *standard CFG*, the origin is given by $\epsilon_t(\varnothing)$, and the guidance direction is $\epsilon_t(c_T) - \epsilon_t(\varnothing)$. In (b) *Adaptive Origin Guidance (ours)*, the origin is interpolated between the identity prediction $\epsilon_t(\langle \texttt{id} \rangle)$ and the standard null prediction, $\epsilon_t(\varnothing)$, as a function of the edit strength. This ensures faithful reconstruction at low strengths while smoothly recovering standard CFG behavior at higher strengths. In (c), we show the edit progression as a function of edit strength. While standard CFG originates from the unconditional prediction (representing arbitrary edits), Adaptive Origin Guidance originates from the identity prediction, creating a trajectory that smoothly transitions from the input image to the target edit.

origin and the standard null origin, as shown in Figure 3b. This design ensures that the edited image remains faithful to the input image at low edit strengths, while recovering the model's standard editing behavior at higher strengths. To obtain this "no-edit" origin, we introduce a dedicated instruction token, $\langle \texttt{id} \rangle$, which corresponds to the identity transformation. We then define the adaptive origin as a function of the edit strength $\alpha \in [0, 1]$:

$$O(\alpha) = s(\alpha)\epsilon(\mathbf{z_t}; c_I, c_T = \varnothing, t) + (1 - s(\alpha))\epsilon(\mathbf{z_t}; c_I, c_T = \langle \texttt{id} \rangle, t),$$

where $s(\alpha)$ is a monotonically increasing scheduler satisfying $s(0) = 0$ and $s(1) = 1$. Integrating this origin into the update step, our final guided prediction becomes:

$$\epsilon^{w,\alpha}(\mathbf{z_t}; c_I, c_T, t) = O(\alpha) + \alpha \cdot w(\epsilon(\mathbf{z_t}; c_I, c_T, t) - \epsilon(\mathbf{z_t}; c_I, \varnothing, t)),$$

where $w$ is set as the standard CFG scale of the model. The resulting guidance term is illustrated in Figures 3b and 3c. We highlight the two boundary conditions of this formulation. First, at $\alpha = 0$, the prediction reduces to $\epsilon(\mathbf{z_t}; c_I, \langle \texttt{id} \rangle, t)$, ensuring the input image remains intact, consistent with a zero-strength edit. Second, at $\alpha = 1$, the equation recovers the standard CFG formulation: $\epsilon(\mathbf{z_t}; c_I, \varnothing, t) + w(\epsilon(\mathbf{z_t}; c_I, c_T, t) - \epsilon(\mathbf{z_t}; c_I, \varnothing, t))$, thereby reproducing the model's default editing behavior.

***Learning the $\langle \texttt{id} \rangle$ Instruction.*** Our formulation relies on the availability of a dedicated instruction that corresponds to the identity editing transformation. To acquire this, we introduce a new token, $\langle \texttt{id} \rangle$, to the text encoder's vocabulary and incorporate it into the training of the editing model. Specifically, we train the model using the standard flow matching objective on a paired editing dataset, which we augment with identity pairs. For these identity samples, we provide identical source and target images paired with the $\langle \texttt{id} \rangle$ instruction, effectively teaching the model to associate this token with faithful reconstruction.

***Implementation Details.*** We employ the Lucy-Edit [DecartAI 2025] architecture as our backbone for both image and video editing, treating images as single-frame videos. To learn the identity instruction, we modify the standard training batch construction by introducing a stochastic mixing strategy. Given a source-target training pair $(I_{src}, I_{tgt})$ and a text instruction $T$, we construct the effective training triplet as follows: with a probability of 10%, we drop the text condition ($T = \varnothing$) to support unconditional prediction; with another 10% probability, we align the $\langle \texttt{id} \rangle$ token with the identity mapping by setting the target equal to the source ($I_{tgt} = I_{src}$) and replacing the text instruction with the token $\langle \texttt{id} \rangle$. In the remaining 80% of cases, we use the standard editing triplet $(I_{src}, I_{tgt}, T)$. We train the full model using this strategy for 3,000 steps following the standard Lucy-Edit protocol. We define the adaptive origin scheduler $s$ as $s(\alpha) = \sqrt{\alpha}$, and ablate this design choice in Section 4.

In Appendix B.2, we further show continuous image editing results with Qwen-Image-Edit [Wu et al. 2025b].

### 3.3 $\langle \texttt{id} \rangle$ Editing Discussion

Previous works have demonstrated that the null prediction in the CFG formulation can be replaced by alternative terms to improve guidance behavior, for example by using a weaker model prediction to increase diversity [Karras et al. 2024]. This raises a question for continuous editing: can we replace the null prediction $\epsilon(\mathbf{z_t}; c_I, \varnothing, t)$ with the $\langle \texttt{id} \rangle$ prediction $\epsilon(\mathbf{z_t}; c_I, \langle \texttt{id} \rangle, t)$, that is, use it both as the guidance origin and in the steering direction? In this section, we show that this seemingly straightforward replacement leads to fundamental instabilities and is therefore unsuitable for continuous editing. In Section 4, we further validate this finding empirically.

We observe that at the final denoising step ($t = 0$), the conditional distribution $p_t(\mathbf{z} \mid c_I, \langle \texttt{id} \rangle)$ collapses to the input image $c_I$, with all probability mass concentrated at that point. Formally, this corresponds to a Dirac delta distribution, $\delta(\mathbf{z} - c_I)$. At earlier timesteps $t > 0$, since the latent is corrupted by Gaussian noise with variance $\sigma_t^2$, this point mass corresponds to a Gaussian distribution centered at $c_I$:

$$p_t(\mathbf{z_t} \mid c_I, \langle \texttt{id} \rangle) \propto \exp\left(-\frac{\|\mathbf{z_t} - c_I\|^2}{2\sigma_t^2}\right).$$

Recall that the noise prediction of a diffusion model is proportional to the scaled score function, $-\sigma_t \nabla_{\mathbf{z_t}} \log p_t(\mathbf{z_t})$. Applying this relation

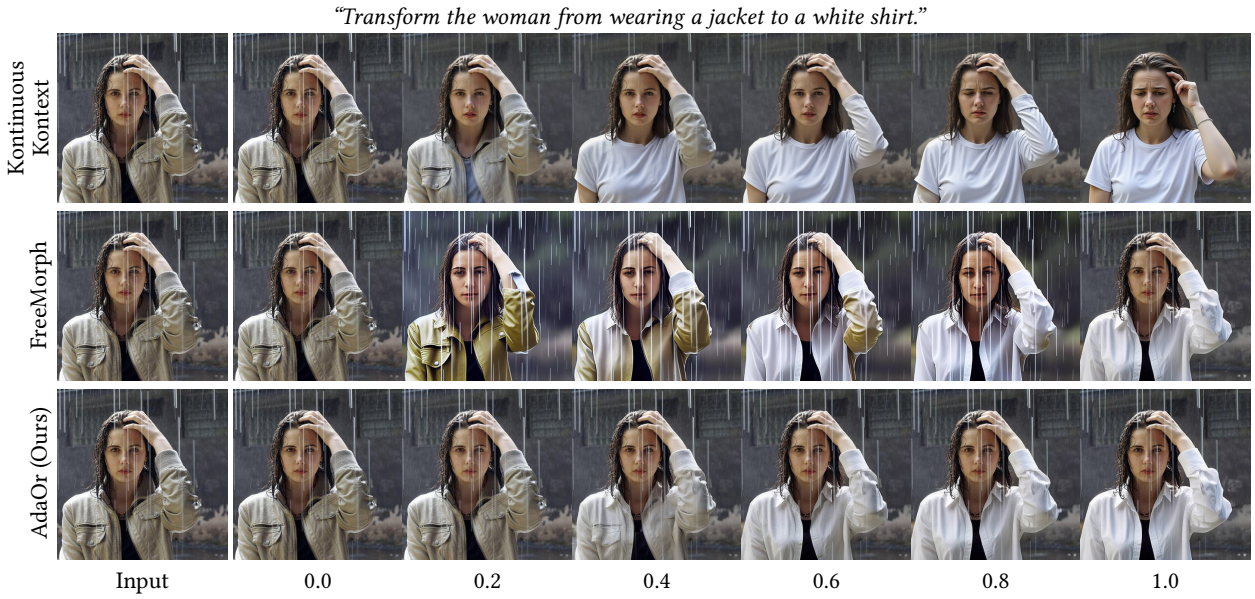



Fig. 4. **Qualitative comparison.** We compare our method (bottom) against Kontinuous Kontext (top) and FreeMorph [Cao et al. 2025] (middle). Kontinuous Kontext [Parihar et al. 2025a] suffers from semantic entanglement, removing the rain and altering the woman's expression as the edit strength increases. FreeMorph fails to generate plausible intermediate states, producing severe artifacts (e.g., the distorted sleeve and hand at strength 0.2). In contrast, AdaOr (Ours) produces a smooth, linear transition that effectively applies the edit while strictly preserving the input content and the subject's appearance.

to the Gaussian distribution induced by $\langle \texttt{id} \rangle$, we obtain:

$$\epsilon(\mathbf{z_t}; c_I, \langle \texttt{id} \rangle, t) \approx \frac{\mathbf{z_t} - c_I}{\sigma_t}.$$

This relationship exposes an inherent instability in the $\langle \texttt{id} \rangle$ prediction. The term corresponds to a score that encourages $\mathbf{z_t}$ to move toward the input $c_I$. This behavior is appropriate when the desired edit strength is low and the latent remains close to the input $c_I$. However, as the edit strength increases and $\mathbf{z_t}$ deviates from the input toward a new concept, the numerator becomes non-zero. As the timestep approaches zero, the noise level $\sigma_t$ vanishes, causing the magnitude of the $\langle \texttt{id} \rangle$ prediction to diverge to infinity. When used in the CFG steering term, this effect is further amplified by the guidance scale, leading to instability near the end of the denoising process for stronger edits.

In contrast, our adaptive origin guidance avoids this behavior by explicitly conditioning on edit strength. When the desired edit strength is low and the latent remains close to the input $c_I$, we rely on the $\langle \texttt{id} \rangle$ term as the guidance origin. As the edit strength increases and $\mathbf{z_t}$ departs from the input, our schedule gradually transitions the origin to the standard null prediction. This null prediction models the broader manifold of natural images and remains well behaved near the end of the denoising process, thereby preventing the guidance explosion.

## 4 Experiments

In this section, we present an extensive evaluation of our method for continuous editing across both video and image domains. Given the dynamic nature of the video results, we encourage readers to refer to the accompanying video and our project page for full visual demonstrations. In the following, we compare our method qualitatively and quantitatively against existing continuous editing baselines and perform ablation studies to validate our design choices.

### 4.1 Comparisons

Since existing continuous editing methods are primarily designed for image generation and lack validation on video models, our analysis focuses on the image domain. We evaluate AdaOr against four baselines representing different continuous editing approaches.

We first compare against FreeMorph [Cao et al. 2025], which morphs between two input images. Since it requires defined endpoints rather than a text instruction, we adapt it for our task by providing the original input image and our edited result at maximum strength (1.0) as the source and target, respectively. Second, we evaluate Kontinuous Kontext [Parihar et al. 2025a], which fine-tunes an image editing model on a synthetic dataset to accept a continuous scalar input for strength control. Next, we compare with Concept Sliders [Gandikota et al. 2023], a method that learns a LoRA-based slider for each edit type. Finally, we compare with SAEdit [Kamenetsky et al. 2025], which leverages Sparse Autoencoders (SAE) trained on human-centered prompts to identify editing directions within the text encoder's latent space. For all baselines, we utilize the official implementations and recommended settings.

***Qualitative Results.*** We present a qualitative comparison with FreeMorph and Kontinuous Kontext in Figure 4, and with Concept Sliders and SAEdit in Figure 5. As shown in Figure 4, FreeMorph fails to generate plausible intermediate images, exhibiting severe structural artifacts such as the distortions in the woman's fingers. Furthermore, the required inversion process degrades the fidelity of the intermediate images. While Kontinuous Kontext yields smoother transitions of higher visual quality, it suffers from semantic entanglement, altering unrelated attributes such as the rain and the subject's


Fig. 5. **Qualitative comparison.** We compare our method (bottom) against Concept Sliders (top) and SAEdit (middle). Concept Sliders modifies the man's identity, while SAEdit introduces only weak curl patterns. AdaOr (Ours) produces a smooth transition that effectively applies the edit while preserving the input content and the subject's identity.

expression. In contrast, our method produces smooth, high-fidelity transitions that strictly adhere to the input image, preserving the original context while effectively applying the requested edit.

As shown in Figure 5, Concept Sliders alters the person's identity, while SAEdit changes the man's hair color but introduces only weak curl patterns. In contrast, AdaOr achieves the desired edit while maintaining a continuous sequence. Additional qualitative results are provided on pages 10–11.

***Quantitative Results.*** Following the evaluation protocol of Kontinuous Kontext, we utilize a subset of PIE-Bench [Ju et al. 2024] to quantitatively evaluate FreeMorph and Kontinuous Kontext. More details in Appendix A.1. Since Concept Sliders and SAEdit are not suitable for the broad scope of PIE-Bench (Concept Sliders requires training a specific LoRA per edit type, and SAEdit was trained on human-centered prompts), we evaluate these methods on the dedicated benchmark provided by SAEdit.

We evaluate our method using $N = 6$ edits with uniform strengths, assessing performance across three dimensions: (i) smoothness: To ensure the editing process is not "jumpy", we measure second-order smoothness using $\delta_{\text{smooth}}$ [Parihar et al. 2025a] and introduce a *Linearity* metric that assesses the uniformity of the editing pace by measuring the variance of perceptual changes between consecutive edit strengths; (ii) text alignment consistency: We introduce *Normalized CLIP-Dir* to verify that every intermediate strength moves semantically towards the target prompt, measuring the average alignment between the local direction at each strength interval and the global text direction; and (iii) perceptual trajectory consistency: To ensure the edit follows a direct path in perceptual space, we measure the cosine similarity between the update at each strength interval and the total edit vector using the DreamSim [Fu et al. 2023] metric. Full mathematical definitions and implementation details are provided in Appendix A.2.

The quantitative results are presented in Tables 1 and 2. As shown in the tables, our method consistently outperforms all baselines

Table 1. **Quantitative evaluation (PIE-Bench).** The top block compares our method to prior work, while the bottom block reports ablations of guidance formulations and scheduling strategies. We evaluate smoothness, text alignment consistency, perceptual trajectory consistency, and linearity.

| Method | $\delta_{\text{smooth}}$ (↓) | Norm. CLIP-Dir (↑) | DreamSim Align (↑) | Linearity (↓) |
|---|---|---|---|---|
| FreeMorph | 0.26 | 1.71 | 0.23 | 0.10 |
| Kontinuous K. | **0.12** | 1.75 | <u>0.32</u> | 0.08 |
| CFG | 0.61 | 1.48 | 0.27 | 0.12 |
| CFG-⟨id⟩ | 0.27 | 1.65 | 0.30 | **0.05** |
| Linear Scheduler | <u>0.14</u> | **1.99** | **0.36** | <u>0.07</u> |
| AdaOr (Ours) | **0.12** | <u>1.89</u> | **0.36** | <u>0.07</u> |

Table 2. **Quantitative evaluation (human-focused).** We compare our method to two other continuous editing methods, evaluating smoothness, text alignment consistency, perceptual trajectory consistency, and linearity.

| Method | $\delta_{\text{smooth}}$ (↓) | Norm. CLIP-Dir (↑) | DreamSim Align (↑) | Linearity (↓) |
|---|---|---|---|---|
| ConceptSliders | 0.26 | 1.82 | 0.35 | 0.15 |
| SAEdit | 0.28 | 2.07 | 0.36 | 0.11 |
| AdaOr (Ours) | **0.24** | **2.21** | **0.37** | **0.10** |

across all evaluated metrics. While Kontinuous Kontext achieves comparable smoothness, our method demonstrates superior text alignment consistency and perceptual trajectory consistency. Furthermore, unlike the competing approaches which are restricted to the image domain, we explicitly demonstrate the extensibility of our method to video editing.

***User Study.*** We further evaluate our method through a user study comparing AdaOr against two strong baselines, Kontinuous Kontext and FreeMorph. The evaluation was conducted on random samples from the PIE-Bench dataset [Ju et al. 2024], using source images and corresponding editing instructions. For AdaOr and Kontinuous Kontext, we generated intermediate images at six strength values ranging from 0 to 1. FreeMorph was evaluated under two settings: one using a Lucy-generated edit at maximum strength to enable a direct comparison of interpolation behavior with AdaOr, and another using a third-party model [Wu et al. 2025a] to assess performance with an alternative endpoint. A total of 36 participants each evaluated 10 randomly assigned tuples, comparing AdaOr's transition sequences against those of a baseline. For each tuple, participants answered three questions: (1) which transition is more linear and smooth, (2) which has more natural-looking intermediate frames, and (3) which result is preferred overall. As shown in Figure 6, AdaOr outperforms both FreeMorph variants across all metrics. Finally, AdaOr achieves comparable intermediate quality and overall preference to Kontinuous Kontext, while outperforming it in transition smoothness.

## 4.2 Ablation Studies

Next, we perform ablation studies. First, we compare against standard CFG by removing the adaptive origin mechanism entirely. Second, we evaluate CFG-⟨id⟩, where we replace the unconditional prediction in CFG with the identity prediction. This is formulated

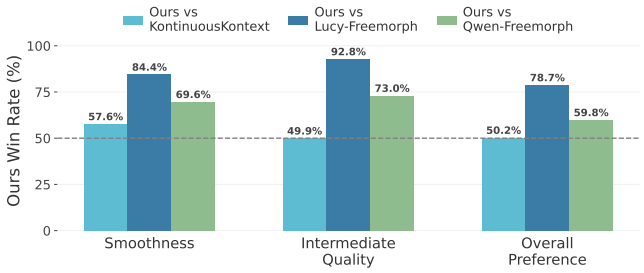



Fig. 6. **User study results.** We report the win rate of our method compared to three baselines: Kontinuous Kontext, Lucy-Freemorph, and Qwen-Freemorph. Human evaluators assessed three aspects: smoothness of the transition, intermediate quality, and overall preference.


Fig. 7. **Qualitative results of the ablation study.** Note that the input image is displayed in the first column of rows 2, 3, and 4. Standard CFG (top) yields arbitrary low-scale content. CFG-⟨id⟩ (row 2) diverges at high scales, while the linear scheduler (row 3) exhibits inconsistent change magnitudes. Our AdaOr (bottom) ensures a smooth, consistent transition.

as: $\epsilon(\mathbf{z_t}; c_I, \langle \texttt{id} \rangle, t) + w(\epsilon(\mathbf{z_t}; c_I, c_T, t) - \epsilon(\mathbf{z_t}; c_I, \langle \texttt{id} \rangle, t))$. Finally, we examine the role of our scheduling strategy by replacing the square root scheduler with a linear scheduler, $s(\alpha) = \alpha$.

We present the qualitative results of the ablation studies in Figure 7. Consistent with previous observations, standard CFG generates arbitrary content at low scales, failing to produce meaningful small-strength edits. As predicted by the analysis in Section 3.3, CFG-⟨id⟩ fails to produce a smooth sequence. As the scale increases, the generation becomes visually exaggerated and distorted, diverging significantly from the target semantics rather than converging to the correct edit. While the linear scheduler produces valid edits, the magnitude of change is inconsistent across the editing strengths. In contrast, our full method successfully balances input preservation with editing fidelity, ensuring a smooth and consistent transition.

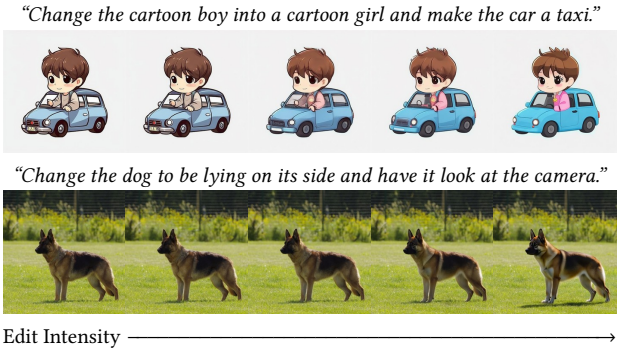



Fig. 8. **Limitations.** We illustrate cases where the backbone model fails to execute the target edit. In the top row, the car fails to transform into a "taxi". In the bottom row, the dog fails to adopt the "lying down" pose. Notably, even in these failure cases, our method generates a smooth and continuous sequence, indicating that the failure stems from the backbone's limited representational scope rather than a failure of our continuity mechanism.

The quantitative ablation results are reported in the lower section of Table 1. As shown, our full method significantly outperforms standard CFG across all metrics. Notably, while CFG-⟨id⟩ achieves a high linearity score, indicating uniform perceptual step sizes, its poor $\delta_{\text{smooth}}$ score reflects a lack of second-order smoothness, consistent with the jagged transitions observed qualitatively. Finally, although the linear scheduler attains performance comparable to our approach, our method exhibits superior smoothness.

### 4.3 Limitations

Our method adopts the training framework of the backbone editing model (e.g., Lucy-Edit). However, as we utilize only a subset of the original training data and a shorter training schedule, our model's representational scope is naturally more constrained. Using our model, we cannot perform continuous edits beyond the backbone's editing capabilities, and we may inherit tendencies toward unnecessary changes or failures in specific edit types. We illustrate two such failure cases in Figure 8. Additionally, our adaptive origin mechanism introduces a slight computational overhead at inference time. While standard CFG requires two noise predictions per step, our method employs three predictions (unconditional, conditional, and ⟨id⟩), resulting in a modest increase in computational cost.

## 5 Conclusions

We presented Adaptive-Origin Guidance (AdaOr), a method for generating smooth, continuous editing sequences from diffusion-based editing models. Our approach is grounded in the observation that at low guidance scales, the generation with CFG is dominated by the unconditional prediction. In the context of editing models, this unconditional prediction corresponds to arbitrary edits rather than an identity edit. By introducing a learnable identity instruction (⟨id⟩), we establish a semantically valid guidance origin, allowing for linear interpolation between the input image and the target edit.

Our method does not require a specialized dataset for continuous edits, nor does it rely on per-edit-type procedures. We demonstrate

that explicitly teaching the model the identity instruction facilitates arithmetic in the prediction space, enabling the execution of complex downstream tasks (e.g., continuous editing). We believe this approach holds further potential for additional generative applications.

For future work, we believe our method can serve as a data generation engine. The continuous sequences produced by our approach could be used to train next-generation editing models that incorporate edit strength directly as a conditioning parameter, distilling inference-time guidance control into native architectural capability.

## Acknowledgments

We would like to express our deep gratitude to Yiftach Edelstein for his invaluable insights and extensive feedback throughout the development of this work. We also thank Yoav Baron, Shelly Golan, Saar Huberman, and Rishubh Parihar for their early feedback and helpful suggestions. Finally, we are grateful to the entire Decart.ai team for their continued support and for providing a stimulating research environment.

## References

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Cross-Image Attention for Zero-Shot Appearance Transfer. arXiv:2311.03335 [cs.CV]

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18208–18218.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. 2025. Stable Flow: Vital Layers for Training-Free Image Editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 7877–7888.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. 2025. Continuous, Subject-Specific Attribute Control in T2I Models by Identifying Semantic Directions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Black Forest Labs. 2024. Flux, https://github.com/black-forest-labs/flux. https://github.com/black-forest-labs/flux

Arwen Bradley and Preetum Nakkiran. 2024. Classifier-Free Guidance is a Predictor-Corrector. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv:2211.09800 [cs.CV]

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. *arXiv preprint arXiv:2304.08465* (2023).

Yukang Cao, Chenyang Si, Jinghao Wang, and Ziwei Liu. 2025. FreeMorph: Tuning-Free Generalized Image Morphing with Diffusion Model. *arXiv preprint arXiv:2507.01953* (2025).

Ta-Ying Cheng, Prafull Sharma, Mark Boss, and Varun Jampani. 2025. MARBLE: Material Recomposition and Blending in CLIP-Space. *CVPR* (2025).

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. 2025. CFG++: Manifold-constrained Classifier Free Guidance for Diffusion Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=E77uvbOTtp

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. 2024. FluxSpace: Disentangled Semantic Editing in Rectified Flow Transformers. arXiv:2412.09611 [cs.CV]

Team DecartAI. 2025. Lucy Edit: Open-Weight Text-Guided Video Editing. (2025).

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:2105.05233 [cs.LG]

Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A. Efros, and Kfir Aberman. 2024. Interpreting the Weight Space of Customized Diffusion Models. arXiv:2406.09413 [cs.CV] https://arxiv.org/abs/2406.09413

Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. 2023. Diffusion Self-Guidance for Controllable Image Generation. (2023).

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. In *Advances in Neural Information Processing Systems*, Vol. 36. 50742–50768.

Rohit Gandikota, Joanna Materzynska, Tingrui Zhou, Antonio Torralba, and David Bau. 2023. Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models. arXiv:2311.12092 [cs.CV] https://arxiv.org/abs/2311.12092

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. 2025. SliderSpace: Decomposing the Visual Capabilities of Diffusion Models. In *Proceedings of the IEEE/CVF international conference on computer vision*. arXiv:2502.01639.

Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. 2025. LoRA-Edit: Controllable First-Frame-Guided Video Editing via Mask-Aware LoRA Fine-Tuning. *arXiv preprint arXiv:2506.10082* (2025).

Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. ReNoise: Real Image Inversion Through Iterative Noising. arXiv:2403.14602 [cs.CV] https://arxiv.org/abs/2403.14602

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2023. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. *arXiv preprint arxiv:2307.10373* (2023).

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. 2026. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233* (2026).

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. 2024. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103* (2024).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626 [cs.CV]

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239 [cs.LG]

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG]

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2023. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. arXiv:2304.06140 [cs.CV]

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. 2025. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 11006–11015.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17191–17202.

Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. 2024. PnP Inversion: Boosting Diffusion-based Editing with 3 Lines of Code. *International Conference on Learning Representations (ICLR)* (2024).

Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. 2025. SAEdit: Token-level control for continuous image editing via Sparse AutoEncoder. *arXiv preprint arXiv:2510.05081* (2025).

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. 2024. Guiding a Diffusion Model with a Bad Version of Itself. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=bg6fVPVs3s

Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. 2024. Noise-free Score Distillation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=dlIMcmlAdk

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

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. 2024. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *arXiv preprint arXiv:2403.14468* (2024).

Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. 2025. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19721–19730.

Black Forest Labs. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742 [cs.GR] https://arxiv.org/abs/2506.15742

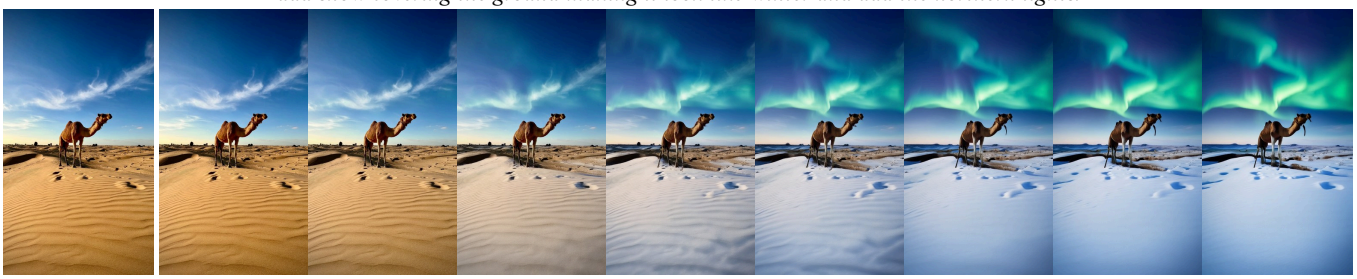
Xiaoming Li, Xinyu Hou, and Chen Change Loy. 2024. When StyleGAN Meets Stable Diffusion: a $\mathcal{W}_+$ Adapter for Personalized Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2187–2196.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. 2025. LightLab: Controlling Light Sources in Images with Diffusion Models. (2025). arXiv:arXiv:2505.09608 doi:10.1145/3721238.3730696

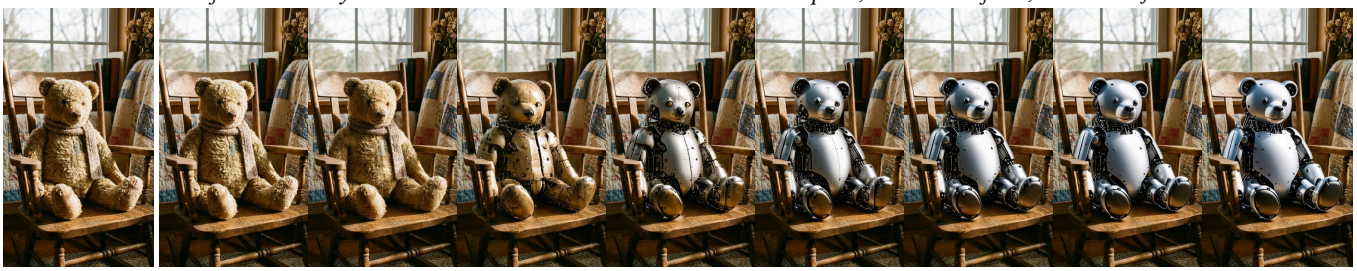
Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic

Differential Equations. arXiv:2108.01073 [cs.CV]
Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. arXiv:2211.09794 [cs.CV]
Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*. PMLR, 16784–16804.
Rishubh Parihar, Or Patashnik, Daniil Ostashev, R Venkatesh Babu, Daniel Cohen-Or, and Kuan-Chieh Wang. 2025a. Kontinuous Kontext: Continuous Strength Control for Instruction-based Image Editing. *arXiv preprint arXiv:2510.08532* (2025).
Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. 2024. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*. Springer, 469–487.
Rishubh Parihar, Sachidanand VS, and R Venkatesh Babu. 2025b. Zero-Shot Depth Aware Image Editing with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15748–15759.
Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot Image-to-Image Translation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23)*. ACM. doi:10.1145/3588432.3591513
Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. Localizing Object-level Shape Variations with Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *arXiv preprint arXiv:2103.17249* (2021).
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]
Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]
L Rout, Y Chen, N Ruiz, C Caramanis, S Shakkottai, and W Chu. 2025. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=Hu0FSOSEyS
Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV]
Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG]
Z-Image Team. 2025. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699* (2025).
Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. 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*. 1921–1930.
Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. arXiv:2503.20314 [cs.CV]
Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025a. Qwen-Image Technical Report. arXiv:2508.02324 [cs.CV] https://arxiv.org/abs/2508.02324
Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025b. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).
Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7623–7633.
Zhenyu Xu, Xiaoqi Shen, Haotian Nan, and Xinyu Zhang. 2025. NumeriKontrol: Adding Numeric Control to Diffusion Transformers for Instruction-based Image Editing. *arXiv preprint arXiv:2511.23105* (2025).
Danah Yatim, Rafail Fridman, Omer Bar-Tal, and Tali Dekel. 2025. DynVFX: Augmenting Real Videos with Dynamic Content. arXiv:2502.03621 [cs.CV] https://arxiv.org/abs/2502.03621
Shai Yehezkel, Omer Dahary, Andrey Voynov, and Daniel Cohen-Or. 2025. Navigating with Annealing Guidance Scale in Diffusion Space. *arXiv preprint arXiv:2506.24108* (2025).
Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. 2024. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7912–7921.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

*"add snow covering the ground making it look like winter and add the northern lights."*



*"Transform the teddy bear into a robotic bear with visible mechanical parts, metallic surfaces, and subtle joints."*



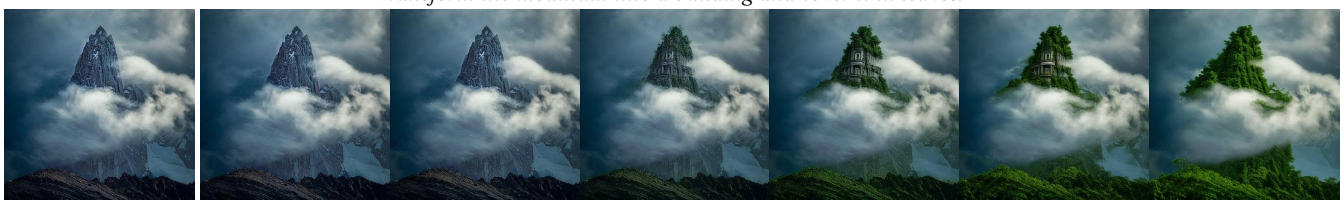
*"the season is autumn the floor is covered with orange leaves."*

*"Transform the horse into a carved wooden horse."*

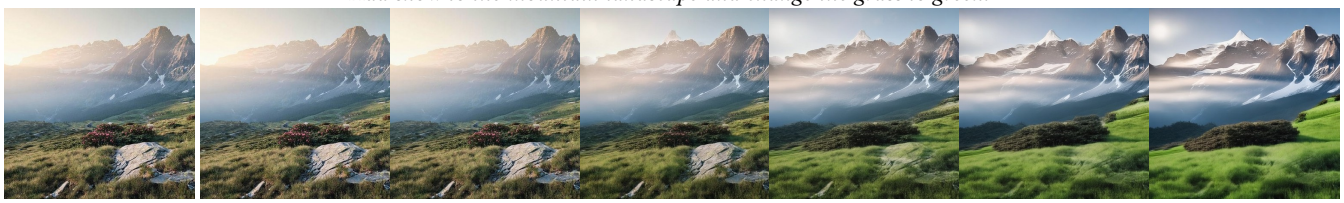
| Input | 0.000 | 0.143 | 0.286 | 0.429 | 0.571 | 0.714 | 0.857 | 1.000 |
|---|---|---|---|---|---|---|---|---|

*"Transform the sea and house into a snowy forest scene."*

*"Transform the mountain into a building and cover it in leaves."*



*"Add snow to the mountain landscape and change the grass to green."*



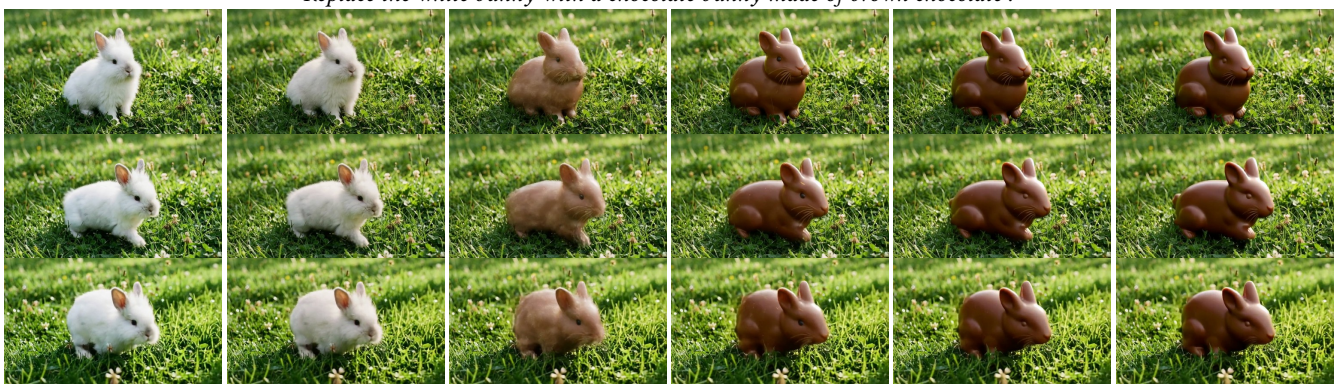
| Input | 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|

*"Add full thick long hair to the man."*

Video Frames

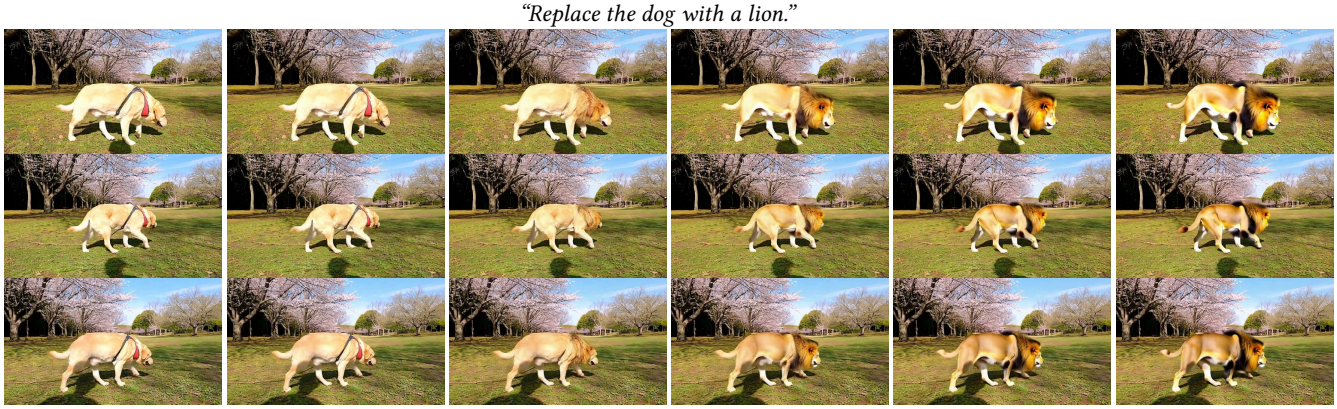
*"Replace the white bunny with a chocolate bunny made of brown chocolate ."*

Video Frames



*"Replace the dog with a lion."*

Video Frames



*"Change the yellow dress to a pink kimono."*

Video Frames

0.000 0.194 0.387 0.613 0.806 1.000

# Appendix

## A Additional Details

### A.1 Benchmarks

***Simplified PIE-Bench.*** Following the protocol of Kontinuous Kontext [Parihar et al. 2025a], we evaluate our continuous editing method on a curated subset of PIE-Bench [Ju et al. 2024]. Specifically, we filter out the roughness, transparency, and style categories, resulting in a dataset of 540 tuples consisting of a source image and an editing instruction. Since PIE-Bench originally contains many samples requiring multiple simultaneous attribute changes, which are less suitable for continuous editing, we employed an LLM to simplify the instructions. Specifically, the LLM was prompted to revise the instructions to target modifications of only one or two attributes in the source image.

***SAEdit Benchmark.*** To compare against SAEdit [Kamenetsky et al. 2025] and Concept Sliders [Gandikota et al. 2023], we adopt the evaluation benchmark proposed in SAEdit. Our benchmark focuses on three specific attribute categories: curly hair, beard, and angry. For each category, an LLM was used to generate pairs of prompts: one describing the source image and the other incorporating the target attribute corresponding to the category. Source images were then generated from the source prompts using FLUX [Black Forest Labs 2024] across multiple random seeds. This benchmark comprises 116 samples in total.

### A.2 Metrics

We generate a sequence of $N = 6$ edits, denoted as $\{I_i\}_{i=1}^{N}$, with editing strengths uniformly distributed between 0 and 1. We assess the performance of each method across the following dimensions:

(1) **Smoothness and Linearity.** We evaluate the smoothness of the edit sequence using the $\delta_{\text{smooth}}$ metric [Parihar et al. 2025a], which measures second-order smoothness. To further quantify the uniformity of the transition, we introduce a *Linearity* metric. We compute the Coefficient of Variation (CV) of the stepwise LPIPS [Zhang et al. 2018] distances. Formally, given the set of distances $S = \{\text{LPIPS}(I_{i+1}, I_i)\}$, we calculate $CV = \sigma(S)/\mu(S)$. Lower values indicate a linear, evenly paced progression, while higher values indicate "jumpy" or irregular transitions.

(2) **Text Alignment Consistency.** We assess whether the semantic change at each step is consistently aligned with the intended text transformation. We introduce the *Normalized CLIP-Dir* metric, which computes the average cosine similarity between the stepwise image direction and the text direction, normalized by the global image-text alignment. This is computed as:

$$\frac{1}{N}\sum_{i=1}^{N-1}\frac{\cos(\Delta\mathbf{v}_{img}^{(i)}, \Delta\mathbf{v}_{text})}{\cos(\Delta\mathbf{v}_{img}^{(global)}, \Delta\mathbf{v}_{text})},$$

where $\Delta\mathbf{v}_{img}^{(i)} = \text{CLIP}(I_{i+1}) - \text{CLIP}(I_i)$ is the local edit direction, $\Delta\mathbf{v}_{img}^{(global)} = \text{CLIP}(I_{N-1}) - \text{CLIP}(I_0)$ is the total edit direction, and $\Delta\mathbf{v}_{text} = \text{CLIP}(c_{\text{edit}}) - \text{CLIP}(c_{\text{src}})$ is the text direction.

(3) **Perceptual Trajectory Consistency.** To ensure the editing process follows a direct perceptual path rather than wandering through arbitrary semantic states, we measure the cosine similarity between the direction of each editing step and the global edit direction in the DreamSim [Fu et al. 2023] embedding space. Formally:

$$\frac{1}{N-1}\sum_{i=0}^{N-2}\cos\left(\text{DS}(I_{i+1}) - \text{DS}(I_i),\quad \text{DS}(I_{N-1}) - \text{DS}(I_0)\right),$$

where $\text{DS}(\cdot)$ denotes the DreamSim [Fu et al. 2023] embedding.

## B Additional Experiments

### B.1 Additional Qualitative Results

We present more qualitative comparison results in Figures 9 to 13.

### B.2 Qwen As Backbone

The main backbone model used in our work for both image and video editing is Lucy-Edit [DecartAI 2025]. However, our method is not specific to the Lucy-Edit architecture and can be readily adapted to other backbone models. We demonstrate this generality by implementing our method using Qwen-Image-Edit [Wu et al. 2025b]. Specifically, we incorporate the $\langle\texttt{id}\rangle$ instruction by training a LoRA module on top of Qwen-Image to learn the identity instruction. All other method details of AdaOr remain unchanged. We present qualitative results in Figure 14.

### B.3 Analytical $\langle\texttt{id}\rangle$ Instruction

In Section 3.3 of the main paper, we demonstrated that the $\langle\texttt{id}\rangle$ prediction can be analytically approximated by:

$$\epsilon(\mathbf{z_t}; c_I, \langle\texttt{id}\rangle, t) \approx \frac{\mathbf{z_t} - c_I}{\sigma_t}.$$

Here, we examine whether this analytical approximation can replace the learned $\langle\texttt{id}\rangle$ prediction employed in AdaOr. Recall that our AdaOr guidance is defined as:

$$\epsilon^{w,\alpha}(\mathbf{z_t}; c_I, c_T, t) = O(\alpha) + \alpha \cdot w(\epsilon(\mathbf{z_t}; c_I, c_T, t) - \epsilon(\mathbf{z_t}; c_I, \varnothing, t)),$$

where the adaptive origin $O(\alpha)$ is:

$$O(\alpha) = s(\alpha)\epsilon(\mathbf{z_t}; c_I, c_T = \varnothing, t) + (1 - s(\alpha))\epsilon(\mathbf{z_t}; c_I, c_T = \langle\texttt{id}\rangle, t).$$

We evaluate the effect of replacing the learned term $\epsilon(\mathbf{z_t}; c_I, c_T = \langle\texttt{id}\rangle, t)$ with the analytical form $(\mathbf{z_t} - c_I)/\sigma_t$ using Lucy-Edit as the backbone. The results are shown in Figure 15, where the first row shows our results, and the second row shows the analytic version.

As observed in the figure, while this substitution produces smooth and continuous sequences, the intermediate outputs lack realism and appear to drift off the natural image manifold. This discrepancy can be attributed to the fundamental difference between Euclidean and manifold interpolation. The analytical prediction essentially pulls the latent linearly toward the input in Euclidean space. In high-dimensional data spaces, such linear paths often traverse low-density regions that do not correspond to valid images. In contrast, the learned $\langle\texttt{id}\rangle$ prediction leverages the model's learned prior, which captures the underlying data distribution. This allows the editing trajectory to traverse the image manifold, ensuring that intermediate steps remain within the space of valid natural images and resulting in significantly more realistic sequences.


Fig. 9. **Qualitative comparison.** We compare our method (bottom) against Kontinuous Kontext (top) and FreeMorph [Cao et al. 2025] (middle).

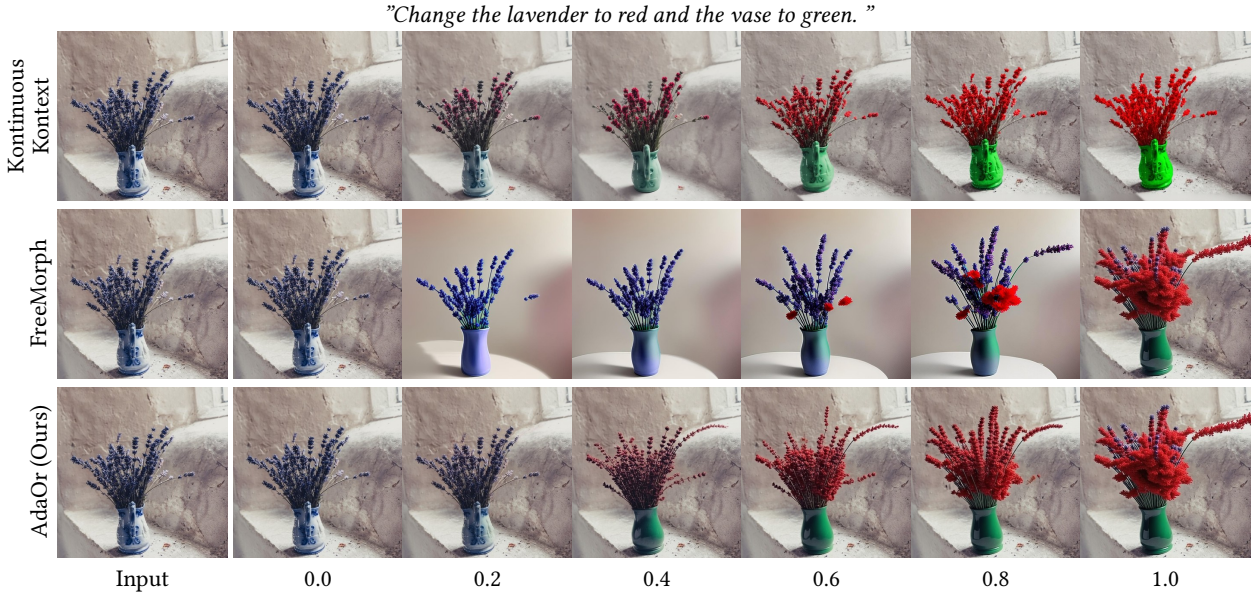



Fig. 10. **Qualitative comparison.** We compare our method (bottom) against Kontinuous Kontext (top) and FreeMorph [Cao et al. 2025] (middle).

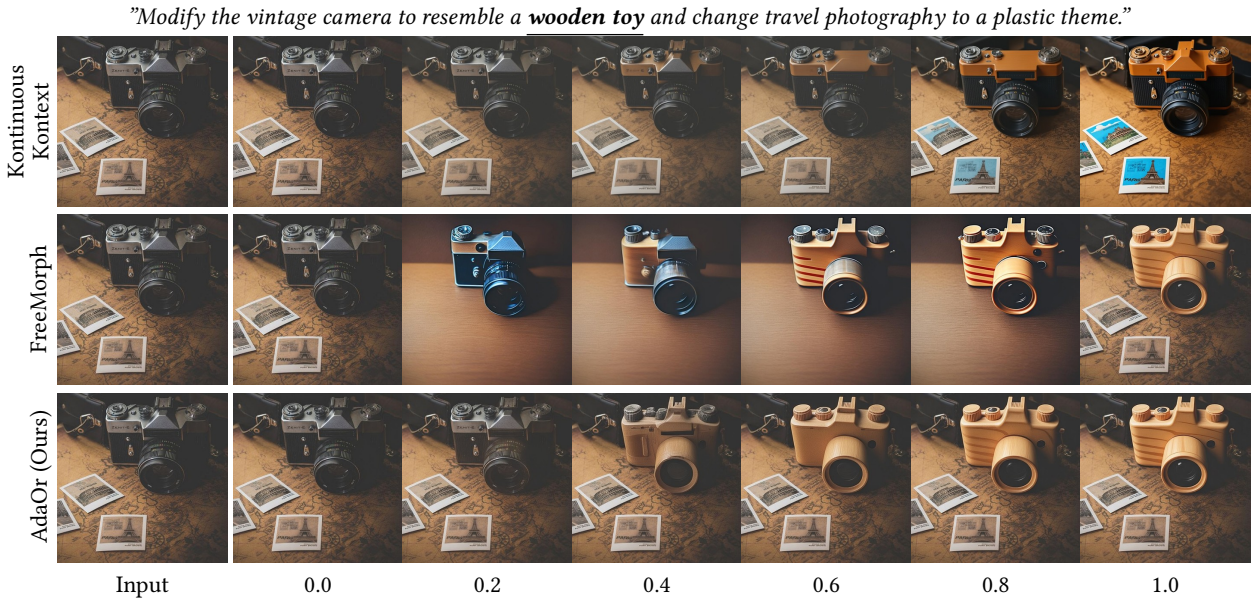



Fig. 11. **Qualitative comparison.** We compare our method (bottom) against Kontinuous Kontext (top) and FreeMorph [Cao et al. 2025] (middle).


Fig. 12. **Qualitative comparison.** Comparison of our method (bottom) against Concept Sliders (top) and SAEdit (middle).

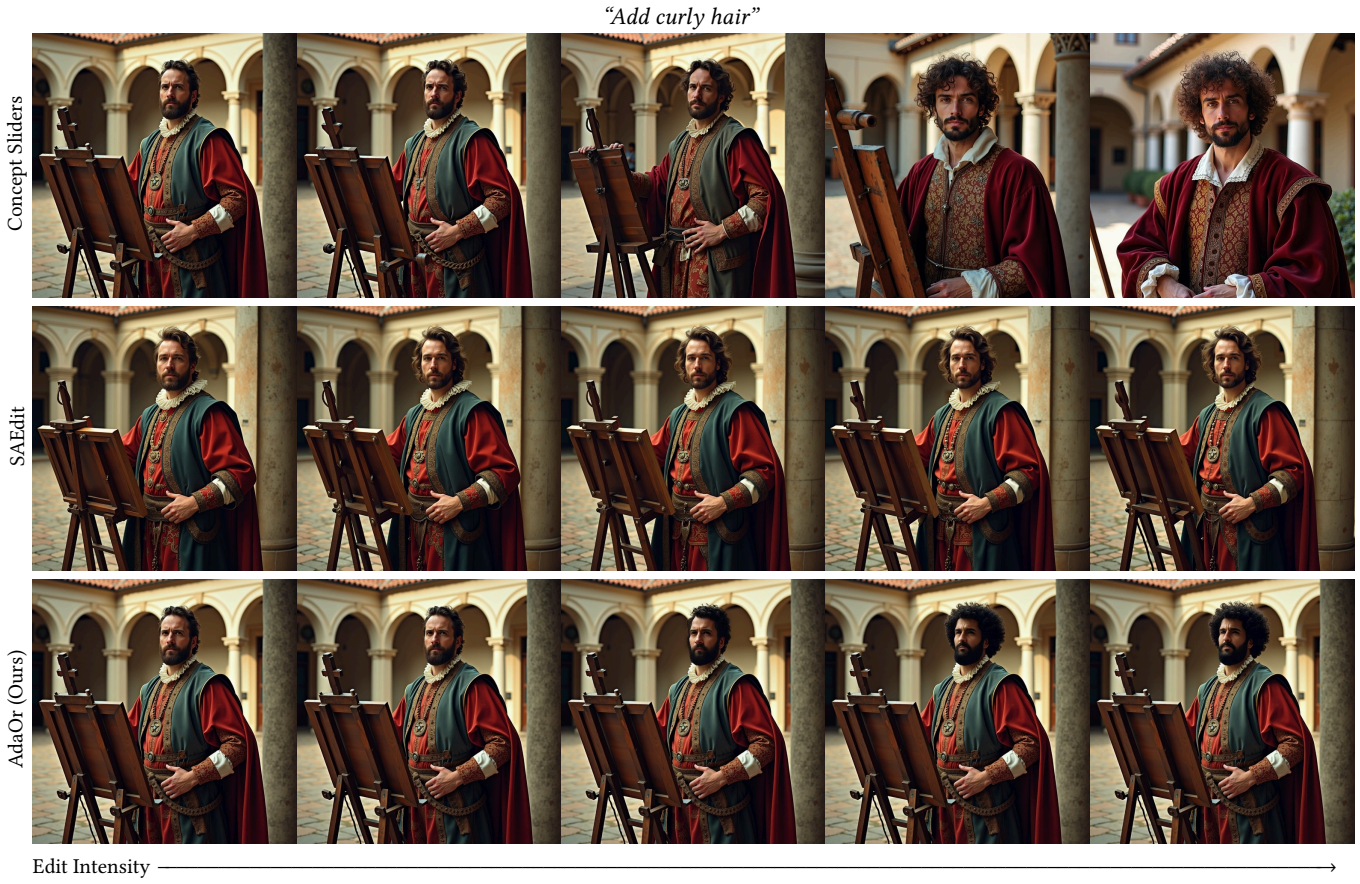



Fig. 13. **Qualitative comparison.** Comparison of our method (bottom) against Concept Sliders (top) and SAEdit (middle).


Fig. 14. **Qualitative results with Qwen-Image-Edit as backbone.**


Fig. 15. **Qualitative results comparing analytical ⟨id⟩ prediction with our learned ⟨id⟩.** Using analytical ⟨id⟩ prediction results in off-manifold intermediate images.