Cora: Correspondence-aware image editing using few step diffusion

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 467

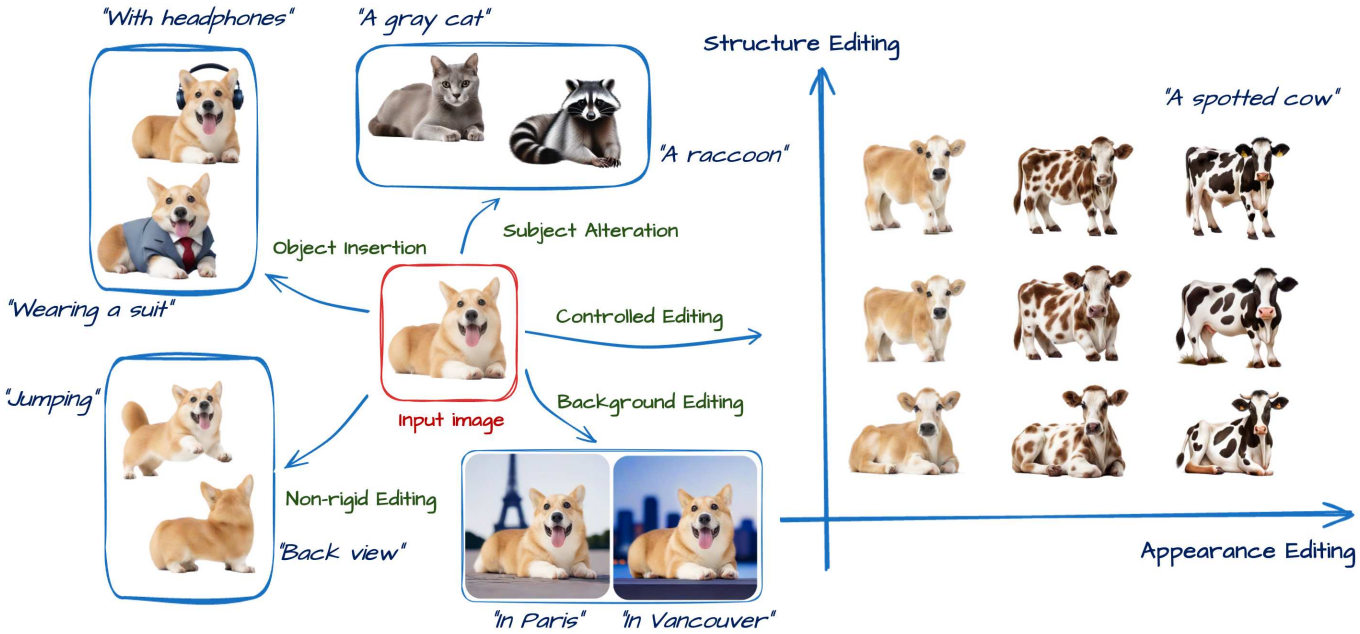


Fig. 1. Cora supports diverse edits, including object insertion, subject and background changes, and non-rigid deformations (e.g., jumping). Our novel correspondence-aware method provides strong control and flexibility for both appearance and structure editing.

Image editing is an important task in computer graphics, vision, and VFX, with recent diffusion-based methods achieving fast and high-quality results. However, edits requiring significant structural changes, such as non-rigid deformations, object modifications, or content generation, remain challenging. Existing few step editing approaches produce artifacts such as irrelevant texture or struggle to preserve key attributes of the source image (e.g., pose). We introduce Cora , a novel editing framework that addresses these limitations by introducing correspondence-aware noise correction and interpolated attention maps. Our method aligns textures and structures between the source and target images through semantic correspondence, enabling accurate texture transfer while generating new content when necessary. Cora offers control over the balance between content generation and preservation. Extensive experiments demonstrate that, quantitatively and qualitatively, Cora excels in maintaining structure, textures, and identity across diverse edits, including pose changes, object addition, and texture refinements. User studies confirm that Cora delivers superior results, outperforming alternatives.

CCS Concepts: • **Generative Models**;

Additional Key Words and Phrases: image editing, diffusion models

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 0730-0301/2025/4-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Anonymous Author(s). 2025. Cora: Correspondence-aware image editing using few step diffusion. *ACM Trans. Graph.* 1, 1 (April 2025), 21 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Image editing is an important task in fields such as computer graphics, computer vision, and VFX. Recent diffusion-based, few-step image editing techniques have significantly improved this process, enabling fast and effective edits with impressive results across diverse scenarios [Deutch et al. 2024; Wu et al. 2024; Xu et al. 2024].

Despite these impressive advancements, edits that require structural changes that go beyond pixel color alteration (e.g., non-rigid edits, object change) remain a challenging task for diffusion models. TurboEdit [Deutch et al. 2024] that relies on noise *correction* to perform edits, often produces artifacts and cannot necessarily preserve identity or important properties of the source image (e.g., pose); see Fig. 2. This is because these corrections do not account for the fact that the generated and source image may not any longer be pixel-aligned after the edit. We resolve this shortcoming by introducing "correspondence-aware" noise corrections that connect source to target pixels by matching their diffusion features.

Edits involving significant deformation of the subject in the image (Fig. 2), often require the generation of new parts, or the exposure of regions not present in the source image. Some approaches

that aim to respect the source for such edits primarily rely on the source image for texture information to maintain the subject's identity [Cao et al. 2023]. While this strategy is somewhat effective, since they inject the intermediate features of the diffusion model from the source image into the self-attention modules [Cao et al. 2023], their edits copy undesired texture from the source into regions of the target image with no clear correspondence (Fig 3:b).

One of our technical contributions is to combine the keys and values that carry texture information from both source and target. This enables the network to generate content when needed while accurately copying textures when relevant information is available in the source image. However, simple methods for combining the source and target, such as concatenation, fail to achieve the desired results (see Sec.4.2). We show that interpolating attention maps enhances performance while offering flexibility and control in both generating new content and preserving existing content.

To achieve the right textures while respecting the structure of the source image, it is also necessary to align attentions by establishing a semantic correspondence. Therefore, we incorporated a correspondence technique (DIFT) into our method whenever source information is available. This technique aligns the attention maps (i.e., keys and values) of the source and target, enabling a more accurate and effective transfer of relevant textures. In the early stages of generation, the model's output is primarily noise, making correspondence infeasible. Therefore, in a four-step diffusion process, we initiate the correspondence process at the last two steps, where the image structure is established, but textures are still being refined. To align the structure of the source and target images, we apply a permutation on queries that are obtained using a matching algorithm. This alignment is performed in the first step of generation, as the image structure takes shape during this phase.

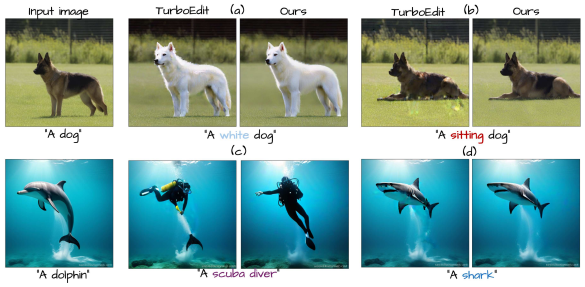


Fig. 2. Comparison between TurboEdit [Deutch et al. 2024] and our correspondence-aware editing approach. Due to misalignment between the source and target images, artifacts are visible in TurboEdit results, such as texture inconsistencies in **(a)**, silhouette artifacts in the legs and fins in **(b, d)**, and undesired elements in **(c)**. Please zoom in for a clearer view of these artifacts.

By combining these strategies, we present Cora, a novel editing method built upon a few-step text-to-image model, SDXL-Turbo [Sauer et al. 2024]. We demonstrate that Cora delivers improved visual results for various edits, thanks to our innovative attention mixing and correspondence-aware techniques. Cora not only excels at preserving the structure of the image and maintaining textures but also supports a wide range of edits, including non-rigid edits (e.g., pose changes), object addition and removal, and texture modifications; see Figure 1. Furthermore, quantitative and qualitative experiments and user studies demonstrate the effectiveness of each component of our method and confirm that Cora surpasses other alternatives.

2 RELATED WORKS

Image editing with diffusion models. Diffusion-based text-to-image models are known for their ability to produce high-quality and diverse outputs [Balaji et al. 2022; Ramesh et al. 2022; Rombach et al. 2021; Saharia et al. 2022]. These models start with Gaussian noise, and iteratively denoise it to generate the final image [Ho et al. 2020; Sohl-Dickstein et al. 2015]. Beyond generation, they can also be leveraged as a powerful prior for visual content editing. This is achieved by adding noise to an image and iteratively denoising it with a new text prompt [Meng et al. 2022], but this method often struggles to balance content preservation with prompt alignment. To address these challenges, follow-up works tweaked the architecture, for example by including editing masks, to better preserve the content in the original image [Avrahami et al. 2022; Hertz et al. 2023a; Mirzaei et al. 2024; Nam et al. 2024; Patashnik et al. 2023; Safaee et al. 2024]. Another approach involves training diffusion models on purpose-built datasets with images, instructional prompts, and ground truth edits [Brooks et al. 2023a], or fine-tuning models on the source image for improved content preservation [Kawar et al. 2023]. However, since training these methods can be both time-consuming and resource-intensive, recent works have focused on leveraging the features of diffusion models for *zero-shot* editing.

Diffusion models' features. The features of diffusion models have been shown to contain rich spatial and semantic information. These features can be used for tasks such as point correspondence [Hedlin et al. 2024, 2023; Luo et al. 2023a; Tang et al. 2023], image and video segmentation [Alimohammadi et al. 2024; Khani et al. 2023], and conditional image generation [Bhat et al. 2023; Luo et al. 2024]. Some approaches leverage the cross-attention maps of diffusion models for localized text-based image editing [Hertz et al. 2022; Mokady et al. 2023]. Tumanyan et al. [2023] inject self-attention and intermediate features of the source image into the generation process of the target image for better content preservation. MasaCtrl [Cao et al. 2023] observes that using the keys and values of the self-attention modules from the source image allows copying its appearance into the target image while enabling non-rigid editing. However, as we will show, only using features from the source image is not sufficient for generating new content in the edited image.

Later works use this observation for style transfer [Alaluf et al. 2023], style-consistent generation [Hertz et al. 2023b], image editing [Koo et al. 2024; Lin et al. 2024; Patashnik et al. 2024] and video editing [Geyer et al. 2023; Qi et al. 2023]. ConsiStory [Tewel et al. 2024] employs attention sharing, and similar to us, employs DIFT-aligned [Tang et al. 2023] feature maps for consistent image generation. Several studies have explored the use of diffusion features for image interpolation, such as Samuel et al. [2024] that investigates various interpolation strategies in the noise space of diffusion models, or He et al. [2024] that examines linear interpolation within the attention space to achieve realistic inbetweening. In Cora, we apply correspondence to the inverted latents of an image to enable

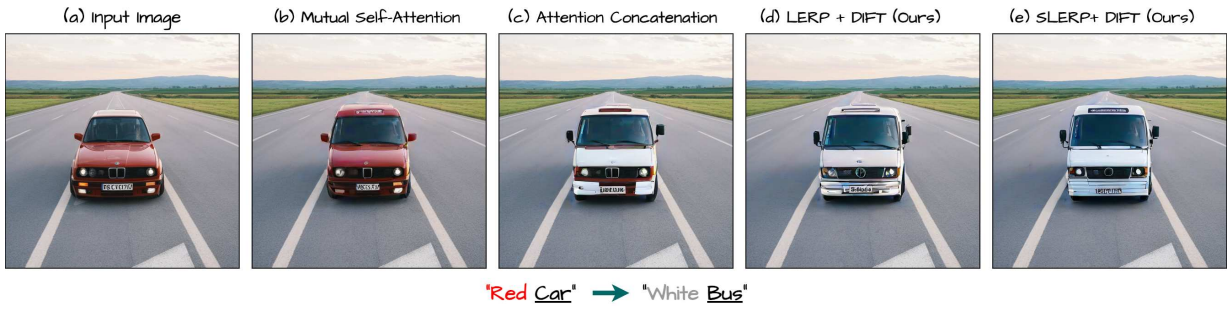


Fig. 3. **Attention mixing strategies. (b):** Using only the source image's keys and values causes artifacts and misalignment with the edit text-prompt. **(c):** Concatenating keys and values leads to undesired appearance bleeding between source and target images. **(d, e):** Aligning and interpolating keys and values produces the best results, with *slerp* providing more realistic and natural outcomes compared to *lerp*.

edits involving significant structural changes. Also, interpolation is performed on *aligned attentions* to balance content preservation and editing flexibility.

Inversion in diffusion models. One way to edit images involves inverting the diffusion model: determining the noise that, when fed to the diffusion model, produces the source image [Song et al. 2022]. Some methods focus on improving this inversion process for more effective editing. DDIM Inversion reverses the deterministic DDIM denoising process [Song et al. 2022], but accumulates small errors during inversion; these errors can lead to significant content drift, especially when using classifier-free guidance [Ho and Salimans 2022] with high guidance [Mokady et al. 2023]. To address this issue, some works rely on iteratively optimizing the intermediate noisy images [Garibi et al. 2024; Li et al. 2024; Pan et al. 2023]. Another line of inversion algorithms [Huberman-Spiegelglas et al. 2024; Wu and la Torre 2023] predicts *both* the initial as well as intermediate noise maps in the DDPM process, enabling better reconstruction and higher editing quality. In our work, we build upon the DDPM inversion proposed by Huberman-Spiegelglas et al. [2024] for its speed, high reconstruction quality, and versatility across editing tasks.

Few step diffusion models. Because of the iterative nature of the diffusion process, these methods are usually slow, requiring 20-100 forward passes of the model. Recent research have focused on designing few-step frameworks using distillation [Salimans and Ho 2022; Yin et al. 2024], consistency constraint [Luo et al. 2023b; Song et al. 2023; Xiao et al. 2023] and adversarial training [Sauer et al. 2024], enabling image generation in 1-8 steps. While these methods are faster, adapting them for editing is not trivial. Several recent works attempted to adapt few-step diffusion to editing tasks. Wu et al. [2024] trains an encoder for fast inversion, Xu et al. [2024] proposes a novel inversion-free framework, and Deutch et al. [2024] adapts editing-friendly DDPM inversion to few-step models. These methods focus on *appearance* changes with minimal structural edits, while the proposed Cora supports both.

3 PRELIMINARIES

Noise-inversion. Huberman-Spiegelglas et al. [2024] maps an input image x_0 to

$$\{x_T, z_{T-1}, \ldots, z_1, z_0\}, \tag{1}$$

where x_T represents the inverted noise at the final timestep T, and z_t denotes correction terms. Given x_0, the algorithm first computes its noisy versions across all timesteps in *forward* (fw) diffusion:

$$x_t^{\text{fw}} = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\tilde{\epsilon}_t, \quad t = 1, \ldots, T, \tag{2}$$

where $\tilde{\epsilon}_t$ is independent standard Gaussian noise, and $\bar{\alpha}_t$ is a parameter of the diffusion scheduler. In the *backward* (bw) process, the scheduler calculates x_{t-1}^{bw} as:

$$x_{t-1}^{\text{bw}} = \mu_t(x_t^{\text{bw}}, c) + \sigma_t z_t, \tag{3}$$

where c is the text prompt used for inversion, and

$$\mu_t(x_t^{\text{bw}}, c) = \frac{1}{\sqrt{\alpha_t}}\left(x_t^{\text{bw}} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}t}}\epsilon_\theta(x_t^{\text{bw}}, t, c)\right) \tag{4}$$

is the predicted x_{t-1} given x_t^{bw}, $\epsilon_\theta(x_t^{\text{bw}}, t, c)$ is the output of the diffusion model, and σ_t is the variance of the scheduler.

At first glance, it is evident that if z_t is standard Gaussian noise, as in the typical denoising process, then x_{t-1}^{bw} and x_{t-1}^{fw} will not necessarily match. However, by setting z_t as:

$$z_t = \frac{x_{t-1}^{\text{fw}} - \mu_t(x_t^{\text{bw}}, c)}{\sigma_t}, \tag{5}$$

we ensure $x_{t-1}^{\text{bw}} \equiv x_{t-1}^{\text{fw}}$. Therefore, using the same text prompt c for generation leads to perfect reconstruction. Modifying the text prompt to a desired edit $\hat{c}$ enables editing, provided the edit does not involve significant structural changes.

TurboEdit. Deutch et al. [2024] introduce a series of modifications to the inversion algorithm to adapt it for the *few-step* setting, including time-shifted inversion, norm-clipping at the final denoising step, and text-guidance to improve prompt alignment. Our method incorporates these modifications.

4 METHOD

We build Cora on top of the few-step diffusion model (i.e., SDXL-Turbo [Sauer et al. 2024]), enabling both appearance editing and structural changes.

Outline. In Section 4.1 we observe that using noise inversion for structural editing introduces artifacts (Fig. 2). We explain how and why a hierarchical correspondence-aware latent correction can resolve these issues. Second, edits that involve significant structural changes or object additions require generating entirely new content or revealing regions absent in the source image. Directly using the keys and values of the source image in the self-attention modules of the diffusion model, as done in MasaCtrl [Cao et al. 2023], often results in unwanted artifacts or misalignment between the target image and appearance editing instructions in the text-prompt (Fig. 3:b). To mitigate this, it is necessary to combine the keys and values of both the source and target images. In Section 4.2, we explore various strategies for this combination and introduce a novel method called *correspondence-aware attention interpolation*. We propose two strategies for interpolation: spherical (SLERP) and linear, and show that SLERP is beneficial. Finally, in Section 4.3, we demonstrate that by matching the queries of the source and target images, we enable control over the extent of structural change in the target image.

4.1 Correspondence-aware latent correction

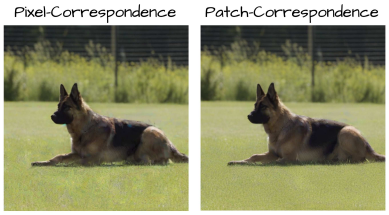
As explained in Section 3, noise-inversion maps the input image x_0 to $\{x_T, z_{T-1}, \ldots, z_1, z_0\}$, where $\{z_t\}$ serve as corrections, ensuring perfect reconstruction when the text prompt c used for inversion *matches* the one used for generation. However, when the text prompt requires a large deformation of the source image, the corrections $\{z_t\}$ are not pixel-aligned with respect to the generated image, leading to severe artifacts (Fig. 2). To address this issue, we align the corrections $\{z_t\}$ with the spatial transformation introduced in the edited image in the final two steps of denoising. This involves constructing a correspondence map $C_{T\to S}$ between the source image I_S and the target image I_T. We achieve this via their DIFT features [Tang et al. 2023], denoted as D_S and D_T:

$$C_{T\to S}(p) = \arg\max_{q\in I_S} \ \mathrm{sim}(D_T(p), D_S(q)), \tag{6}$$

where p and q are pixels from the target and source images, respectively, and sim is the cosine similarity between the two feature vectors. We construct an aligned correction term z_t^{aln} as:

$$z_t^{\mathrm{aln}}(p) = z_t(C_{T\to S}(p)). \tag{7}$$

Patch correspondence Since DIFT features might be noisy and inaccurate, the resulting target image often contains artifacts(inset -left). Therefore, we propose a patch-wise correspondence approach. The DIFT features D_S and D_T are divided into small, overlapping patches, and correspondence is computed for each patch rather than individual pixels. For each patch, we concatenate its pixel-wise DIFT features and calculate cosine similarity between these concatenated features. Due to overlapping patches, multiple patches may contribute to the alignment of a single pixel p. To obtain the final aligned correction latent $z_t^{\mathrm{aln}}(p)$, we average the contributions from all overlapping patches at pixel p.



As denoising progresses and the features become less noisy at later timesteps, the size of the patches is gradually reduced. This ensures that the correspondence becomes more precise, adapting dynamically to the evolving reliability of the features (see inset -right).

4.2 Correspondence-aware attention interpolation

Achieving high-quality image editing requires balancing the preservation of key aspects of the source image (e.g., appearance or identity) with the introduction of new elements or modifications. Recent approaches often achieve this by injecting the attention features of the source image into the denoising process of the target image [Cao et al. 2023]. While this method is effective, it overlooks the fact that editing often involves generating *new* content, and this content may lack clear correspondence to content stored within the source image. We now consider existing methods, and present a novel strategy for combining attentions between source and target images.

Mutual self-attention. MasaCtrl [Cao et al. 2023] uses the source image's keys and values in the self-attention modules of the diffusion model. This ensures that the context of the source image, such as its appearance and identity are preserved:

$$f^e = \mathrm{Attention}(Q_T, K_S, V_S), \tag{8}$$

where Q_T is the query of the target image, K_S and V_S are the keys and values of the source image, and f^e represents the output of the self-attention module. While this strategy effectively retains the appearance and identity of the original content, it also limits the model's ability to generate *new* content, such as adding objects or significantly altering appearances (see Fig. 3:b).

Concatenation. An approach to incorporate appearance editing is concatenating the keys and values of the source and target images as done by Hertz et al. [2023b], Deng et al. [2023], and Tewel et al. [2024]:

$$f^e = \mathrm{Attention}\big(Q_T, [\lambda \cdot K_S, K_T], [\lambda \cdot V_S, V_T]\big), \tag{9}$$

where [,] denotes concatenation, and λ is a scaling factor that controls how much the source appearance affects the target image. While this enables appearance changes, it often fails to achieve a smooth interpolation between the two appearances. This can result in unnatural "appearance leakage", such as elements of the *red car* blending into the *white bus* (see Fig. 3:c).

Linear interpolation. Another approach is to linearly interpolate the keys and values of the source and target images [He et al. 2024]. Differently from He et al. [2024], we linearly interpolate after matching features between source and target features (Sec. 4.1), as otherwise this may cause artifacts due to the mis-alignment of source

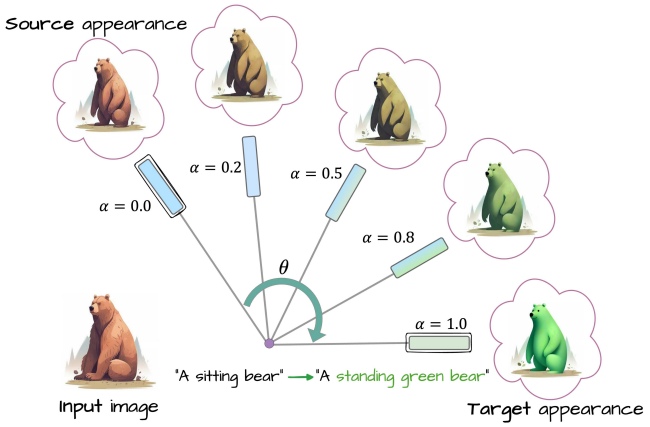


Fig. 4. **Adjusting** α Provides control over the appearance transition between the source image and the target appearance. When $\alpha = 0$, the appearance is entirely derived from the source image, and when $\alpha = 1$, it fully reflects the editing prompt. Intermediate values of α allow for a gradual blend, enabling fine-grained control between these two extremes.

and target (see the inset):

$$\mathcal{L}(v_1, v_2, \alpha) = (1-\alpha)\cdot v_1 + \alpha \cdot v_2, \tag{10}$$

$$f^e = \text{Attention}(Q_T, \mathcal{L}(K_S^{\text{aln}}, K_T, \alpha), \mathcal{L}(V_S^{\text{aln}}, V_T, \alpha)). \tag{11}$$

While this approach is somewhat effective, it sometimes causes unwanted artifacts when interpolating between features that are significantly different (see the mirrors in Figure 3:d). To address this limitation, we explore the use of *spherical linear interpolation (SLERP)* for interpolating between the keys and values, which takes vector directions into account for smoother blending:

$$\mathcal{Q}(v_1, v_2; \alpha) = \frac{\sin((1-\alpha)\theta)}{\sin(\theta)} \cdot \frac{v_1}{|v_1|} + \frac{\sin(\alpha\theta)}{\sin(\theta)} \cdot \frac{v_2}{|v_2|}, \tag{12}$$

where θ is the angle between the vectors v_1 and v_2, and $\alpha \in [0, 1]$ is the blending weight. The output of this operation is a unit vector. To preserve the magnitude information of the vectors, we also interpolate their magnitudes as $\mathcal{M}(v_1, v_2; \alpha) = (1-\alpha)\cdot|v_1| + \alpha \cdot |v_2|$. Combining these, the full interpolation formula becomes:

$$\mathcal{MQ}(v_1, v_2; \alpha) = \mathcal{M}(v_1, v_2; \alpha) \cdot \mathcal{Q}(v_1, v_2; \alpha). \tag{13}$$

In the self-attention modules, this results in:

$$f^e = \text{Attention}\Big(Q_t,\ \mathcal{MQ}(K_S^{\text{aln}}, K_T; \alpha),\ \mathcal{MQ}(V_S^{\text{aln}}, V_T; \alpha)\Big). \tag{14}$$

This ensures that transitions between source and target vectors respect their angular relationships and provides smoother and more reliable blending between the appearance of the source and target images (see Fig. 3:e). Furthermore, adjusting α enables *control* over the extent of appearance changes in the target image relative to the source image (see Fig. 4).

Content-adaptive interpolation. In situations where the prompt suggests the insertion of a new object or a deformation that causes significant disocclusion, expecting that every generated/target pixel matches a pixel in the source image is not reasonable. In such scenarios, naively aligning the keys and values of the source and target

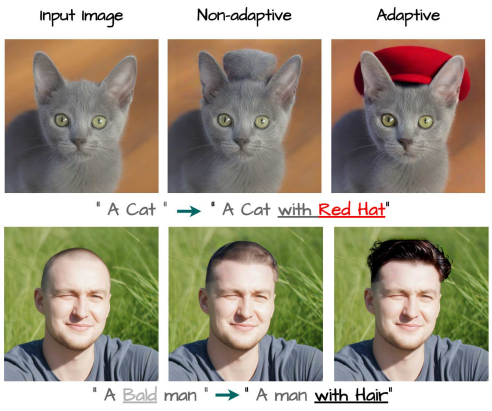


Fig. 5. **Effect of content-adaptive interpolation.** Interpolating the keys and values for target image regions without clear correspondence in the source image results in undesirable edits (**b**). Classifying these regions and using only the target keys and values mitigates this issue (**c**).

images leads to text misalignment and/or visual artifacts (see Fig. 5). To address this issue, we propose to *classify* pixels in the target image that do not have any correspondence in the source image (i.e. "new" pixels). We achieve this by a *bidirectional* comparison between source and target image patches. Specifically, for each patch in the source image $s \in \mathcal{S}$, we compute the set $K(s)$ of k-nearest target patches:

$$K(s) = \arg\operatorname*{top}_{k,\ t\in\mathcal{T}} \ \text{sim}(s, t), \tag{15}$$

where $\text{sim}(s, t)$ is the cosine similarity from Section 4.1. Similarly, for each patch in the target image $t \in \mathcal{T}$, we define its top-k nearest patches as $K(t)$. We call a pair of patches (s, t) *bidirectionally matched* if $t \in K(s)$ and $s \in K(t)$. For bidirectionally *matched* patches, we use the hyper-parameter α specified by the user, while by setting $\alpha = 1$ we could let patches being purely driven by the text, rather than the source image. In particular, for *unmatched* target patches $t \in \mathcal{U}$ we set $\alpha = 1$ if we deem the correspondence to be particularly "weak". We measure this weakness by computing the score of the best available match:

$$\text{sim}_{max}(t) = \max_{s\in\mathcal{S}} \ \text{sim}(s, t) \tag{16}$$

and determining which subset of $t \in \mathcal{U}$ have the worst matches by computing the γ-quantile of the scores in this set:

$$\tau_\gamma = \text{quantile}_\gamma\big(\{\text{sim}_{\max}(t) \mid t \in \mathcal{U}\}\big). \tag{17}$$

finally setting α=1 whenever $sim_{max}(t) < \tau_\gamma$. We typically set γ=3% in our experiments.

4.3 Structural alignment

Preserving the overall layout of the image (i.e. preserving structure) is important when editing images. Recent works [Alaluf et al. 2023; Cao et al. 2023] have demonstrated that it is the queries in the self-attention modules of diffusion models that specify the structure of the generated image. Hence, while in Section 4.2 we described key/value mixing, we now incorporate *queries* from the source image during the denoising process to retain the overall image structure. Our key idea is that reproducing the structure of the original image,

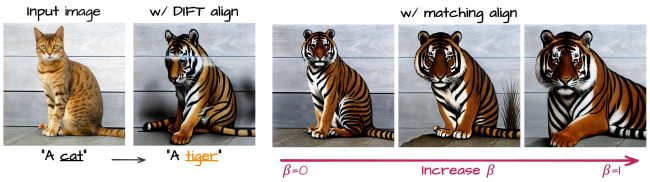


Fig. 6. **Structure Alignment.** DIFT-aligned queries produce unnatural edits, while our matching algorithm with adjustable blending weights (β) enables transitions between full structure alignment and new layout generation.

up to a non-rigid deformation, implies that we want to find *all* the local structures of the source image within the generated target. We achieve this via Hungarian matching [Kuhn 1955] between source and target queries, as this provides us with a *one-to-one* matching (i.e. every target query should match one source query). In particular, Hungarian matching computes the optimal permutation given a weight matrix C, which then shuffles our generation queries:

$$\pi^* = \underset{\pi \in \mathrm{Perm}(N)}{\arg\min} \sum_{n=1}^{N} C[n, \pi(n)], \quad Q_T^{\pi}[n] \leftarrow Q_T[\pi^*(n)] \tag{18}$$

We define our weight matrix C as a linear interpolation controlled by a blending weight β between two matrices (described next):

$$C = (1-\beta) \cdot \mathrm{normalize}(C_{\mathrm{SA}}) + \beta \cdot \mathrm{normalize}(C_{\mathrm{TC}}), \tag{19}$$

where normalize(.) rescales the matrices to ensure comparability. The first matrix C_{SA} encourages the target queries to remain similar to the source queries, hence promoting *Source Alignment*:

$$C_{\mathrm{SA}}[i,j] = 1 - \mathrm{sim}(q_s[i], q_t[j]). \tag{20}$$

The second matrix C_{TC} attempts to penalize index differences among the target queries, hence promoting *Target Consistency*:

$$C_{\mathrm{TC}}[i,j] = \sqrt{|i-j|}. \tag{21}$$

Varying β provides us with control over the structure of the target image, transitioning between preserving the source structure ($\beta \approx 0$) or adhering more to the text prompt while remaining self-consistent ($\beta \approx 1$). An illustration of the effect of β can be found in Figure 6. Note that this process is limited to the first step of denoising when the coarse structure of the generated image is established.

5 EXPERIMENTS

In this section, we present various editing results on real images, demonstrating the versatility of our method. We then compare our approach with both multi-step and few-step baselines to highlight its advantages in terms of visual quality and speed. Finally, we conduct ablation studies to analyze the contribution of each component.

Qualitative Results. Fig. 8 showcases several edits generated by our 4-step denoising procedure. These examples include non-rigid deformations, inserting new objects, and replacing existing objects. Our method generally preserves the overall structure of the input image while accurately reflecting the requested edits.

To compare with existing approaches, Figure 9 illustrates that our approach is more successful at maintaining the subject's identity and reducing artifacts. We focus on few-step baselines such as

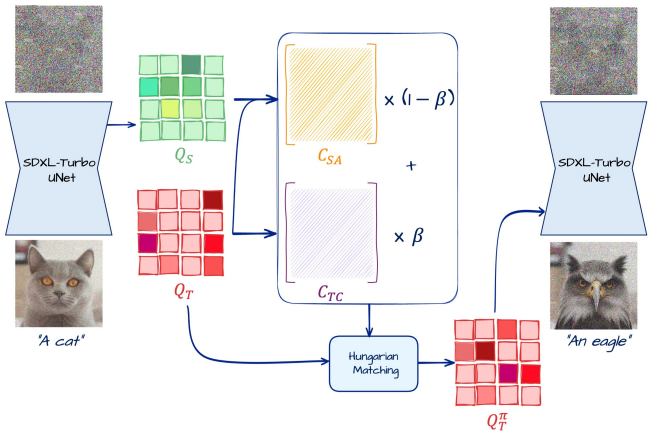


Fig. 7. **Structural Alignment.** In the first denoising step, we extract self-attention queries from both source and target images. We then define two cost matrices: C_{SA}, which promotes structural alignment between source and target, and C_{TC}, which preserves target structure. By linearly combining these matrices, we can control the strength of alignment. The resulting cost matrix is then used in the Hungarian matching algorithm to permute the target queries, aligning them with the source's structure.

TurboEdit [Wu et al. 2024] and InfEdit [Xu et al. 2024] as they operate in a similar few-step regime, as well as multi-step frameworks such as MasaCtrl [Cao et al. 2023] and Edit friendly DDPM inversion [Huberman-Spiegelglas et al. 2024]. Our results exhibit fewer distortions and better fidelity, particularly upon closer inspection.

We further expand the comparison to multi-step methods, including Prompt-to-Prompt (P2P) [Hertz et al. 2022], plug-and-play (PnP) [Tumanyan et al. 2022], instruct-pix2pix [Brooks et al. 2023b], and StyleDiffusion [Li et al. 2023] (see Figure 10). Despite being significantly faster, our few-step approach achieves comparable or superior results in preserving details and adhering to the edits.

User studies. Although we quantitatively show in Supplementary that across different metrics Cora is comparable or superior to alternatives, standard metrics (e.g., PSNR and LPIPS) often fail to capture the visual quality of edits with significant deformation—the key focus of our paper. Therefore, we also conducted a user study. Participants were shown the original image, the editing prompt, and outputs from our method and various baselines. They ranked the images based on (i) alignment with the prompt and (ii) preservation of the subject in the source image, using a scale from 1 (worst) to 4 (best). The average rankings are summarized in Tab.1. Feedback from 51 participants strongly favored our method over other few-step approaches and found it comparable to computationally intensive multi-step techniques. Additionally, a separate user study on attention mixing strategies revealed that correspondence-aligned SLERP interpolation produced the best results, as shown in Tab.2.

Ablation Studies. We now examine the contributions of the main components in our framework:

Structure Alignment. Disabling structure alignment reduces background fidelity,although the edited object remains well-aligned to the text prompt. Visual comparisons (see Supplementary) confirm that structure alignment is crucial for preserving scene details.

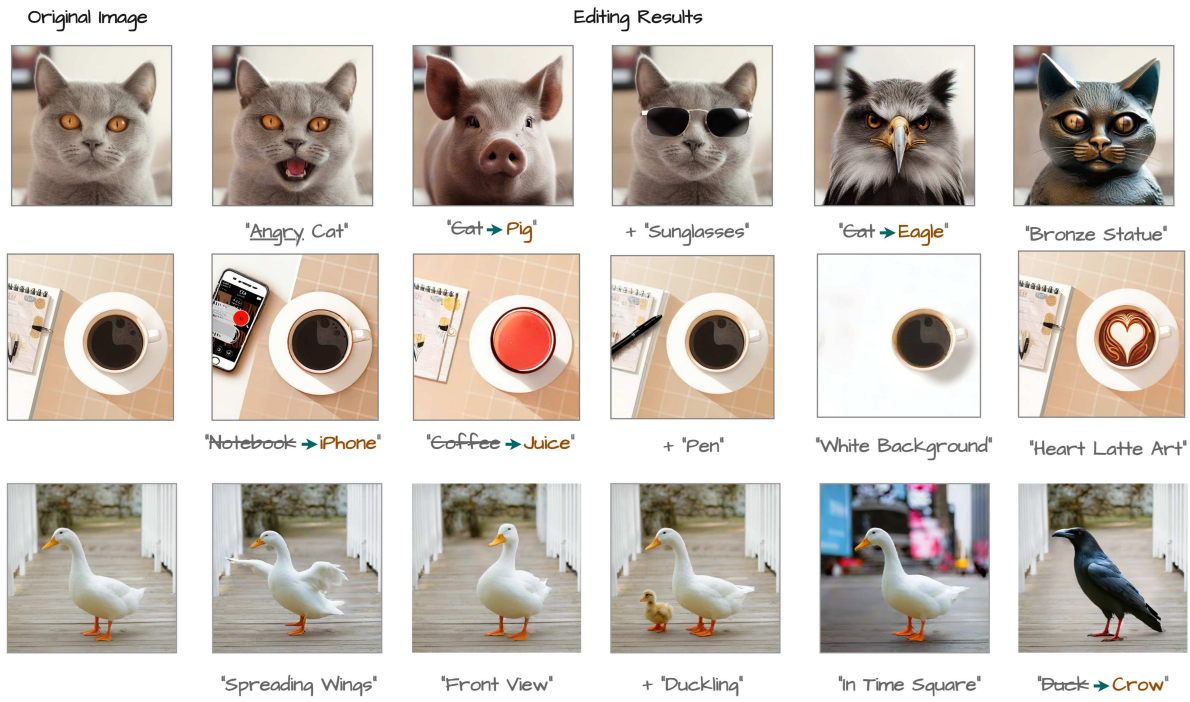


Fig. 8. **Qualitative results**. We demonstrate the ability of our method to perform various types of edits on multiple images.

Table 1. **User study**. Our method has received a significantly higher score than the alternatives.

Method	Average Ranking (↑)
MasaCtrl [Cao et al. 2023]	1.02
DDPM Inversion [Huberman-Spiegelglas et al. 2024]	1.78
InfEdit [Xu et al. 2024]	1.67
TurboEdit [Deutch et al. 2024]	2.24
Ours	**3.29**

Table 2. **User study**. Ablation user study on different attention mixing strategies mentioned in Sec. 4.2. It is evident that correspondence-aligned SLERP interpolation yields the best results.

Method	Average Ranking (↑)
Mutual	0.40
Concatenation	1.71
LERP	2.08
LERP+DIFT	2.58
SLERP+DIFT	**3.23**

Correspondence-Aware Latent Correction. Removing this module causes significant distortions in the edited region. Hence, the latent correction is essential for producing coherent edits. For visual results, see Supplementary.

SLERP vs. LERP. While switching from SLERP to LERP often produces similar outcomes, SLERP can yield more consistent transitions in certain challenging cases. For visual results, see Supplementary.

Removing correspondence alignment From Attention. As seen in Sec. 4.2, this leads to more artifacts, as it helps enforce alignment between the modified content and the background.

6 CONCLUSION, LIMITATIONS, FUTURE WORK

We have introduced Cora, a novel diffusion-based image editing method that addresses the challenges of structural edits, such as non-rigid transformations and object insertions. By leveraging innovative attention mixing and correspondence-aware techniques, our approach enables accurate texture preservation and structural alignment. Unlike existing methods, Cora effectively generates new content when required while maintaining fidelity to the source image where relevant. Our results demonstrate significant improvements in visual quality and flexibility across a wide range of editing tasks, including pose alterations, object manipulation, and texture modifications. Quantitative evaluations further validate that Cora consistently outperforms state-of-the-art methods in both quality and versatility. However, our method still suffers from some limitations. For example, prompts may change unintended parts of the image (e.g., changing a car's color might also affect the background). This can be resolved by using masks automatically obtained via cross and self-attention. While this is a promising direction, it is challenging with only four steps denoising and can be considered as a future work. Another future directions could be to extend Cora to video editing or evaluate alternative non-linear interpolation techniques for attentions.

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]

Amirhossein Alimohammadi, Sauradip Nag, Saeid Asgari Taghanaki, Andrea Tagliasacchi, Ghassan Hamarneh, and Ali Mahdavi Amiri. 2024. SMITE: Segment Me In TimE. arXiv:2410.18538 [cs.CV] https://arxiv.org/abs/2410.18538

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 (CVPR)*. 18208–18218.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. 2022. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324* (2022).

Shariq Farooq Bhat, Niloy J. Mitra, and Peter Wonka. 2023. LooseControl: Lifting ControlNet for Generalized Depth Conditioning. arXiv:2312.03079 [cs.CV]

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023a. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

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

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22560–22570.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. 2023. Z^*: Zero-shot Style Transfer via Attention Rearrangement. arXiv:2311.16491 [cs.CV] https://arxiv.org/abs/2311.16491

Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. 2024. TurboEdit: Text-Based Image Editing Using Few-Step Diffusion Models. arXiv:2408.00735 [cs.CV] https://arxiv.org/abs/2408.00735

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]

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).

Qiyuan He, Jinghao Wang, Ziwei Liu, and Angela Yao. 2024. AID: Attention Interpolation of Text-to-Image Diffusion. *arXiv preprint arXiv:2403.17924* (2024).

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. 2024. Unsupervised Keypoints from Pretrained Diffusion Models.

Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 2023. Unsupervised Semantic Correspondence Using Stable Diffusion.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. 2023a. Delta Denoising Score. (2023).

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. (2022).

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2023b. Style Aligned Image Generation via Shared Attention. (2023).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG] https://arxiv.org/abs/2207.12598

Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2024. An Edit Friendly DDPM Noise Space: Inversion and Manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *Conference on Computer Vision and Pattern Recognition 2023*.

Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. 2023. SLiMe: Segment Like Me. arXiv:2309.03179 [cs.CV]

Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D Yoo. 2024. FlexiEdit: Frequency-Aware Latent Refinement for Enhanced Non-Rigid Editing. *arXiv preprint arXiv:2407.17850* (2024).

H. W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly* 2, 1-2 (1955), 83–97. https://doi.org/10.1002/nav.3800020109 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800020109

Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. 2024. Source Prompt Disentangled Inversion for Boosting Image Editability with Diffusion Models. (2024).

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. StyleDiffusion: Prompt-Embedding Inversion for Text-Based Editing. *arXiv preprint arXiv:2303.15649* (2023).

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. 2024. Ctrl-X: Controlling Structure and Appearance for Text-To-Image Generation Without Guidance. In *Advances in Neural Information Processing Systems*.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. 2024. Readout Guidance: Learning Control from Diffusion Features. *CVPR*.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023a. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. In *Advances in Neural Information Processing Systems*.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023b. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv:2310.04378 [cs.CV]

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. In *International Conference on Learning Representations*.

Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G. Derpanis, and Igor Gilitschenski. 2024. Watch Your Steps: Local Image and Scene Editing by Text Instructions. In *ECCV*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text Inversion for Editing Real Images using Guided Diffusion Models. (2023).

Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. 2024. Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9192–9201.

Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, and Stephen Huang. 2023. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15912–15921.

Or Patashnik, Rinon Gal, Daniel Cohen-Or, Jun-Yan Zhu, and Fernando De La Torre. 2024. Consolidating Attention Features for Multi-view Image Editing. In *SIGGRAPH Asia 2024 Conference Papers (SA '24)*. Association for Computing Machinery, New York, NY, USA, Article 40, 12 pages. https://doi.org/10.1145/3680528.3687611

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)*.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. FateZero: Fusing Attentions for Zero-shot Text-based Video Editing. *arXiv:2303.09535* (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV] https://arxiv.org/abs/2103.00020

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical textconditional image generation with clip latents. arXiv:2204.06125 [cs.CV]

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752* (2021). arXiv:2112.10752

Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. 2024. CLiC: Concept Learning in Context. *CVPR* (2024).

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. https://doi.org/10.48550/ARXIV.2205.11487

Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=TIdIXIpzhoI

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. 2024. Norm-guided latent space exploration for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2522, 13 pages.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. 2024. Fast High-Resolution Image Synthesis with Latent Adversarial Diffusion Distillation. arXiv:2403.12015 [cs.CV] https://arxiv.org/abs/2403.12015

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 2256–2265. https://proceedings.mlr.press/v37/sohl-dickstein15.html

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG] https://arxiv.org/abs/2010.02502

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*. 32211–32252.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent Correspondence from Image Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=ypOiXjdfnU

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-Free Consistent Text-to-Image Generation. arXiv:2402.03286 [cs.CV] https://arxiv.org/abs/2402.03286

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2022. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. arXiv:2211.12572 [cs.CV] https://arxiv.org/abs/2211.12572

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 (CVPR)*. 1921–1930.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Chen Henry Wu and Fernando De la Torre. 2023. A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance. In *ICCV*.

Zongze Wu, Nicholas Kolkin, Jonathan Brandt, Richard Zhang, and Eli Shechtman. 2024. TurboEdit: Instant text-based image editing. *ECCV* (2024).

Jie Xiao, Kai Zhu, Han Zhang, Zhiheng Liu, Yujun Shen, Yu Liu, Xueyang Fu, and Zheng-Jun Zha. 2023. CCM: Adding Conditional Controls to Text-to-Image Consistency Models. *arXiv preprint arXiv:2312.06971* (2023).

Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. 2024. Inversion-Free Image Editing with Natural Language. (2024).

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. 2024. One-step Diffusion with Distribution Matching Distillation. *CVPR* (2024).

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924 [cs.CV] https://arxiv.org/abs/1801.03924

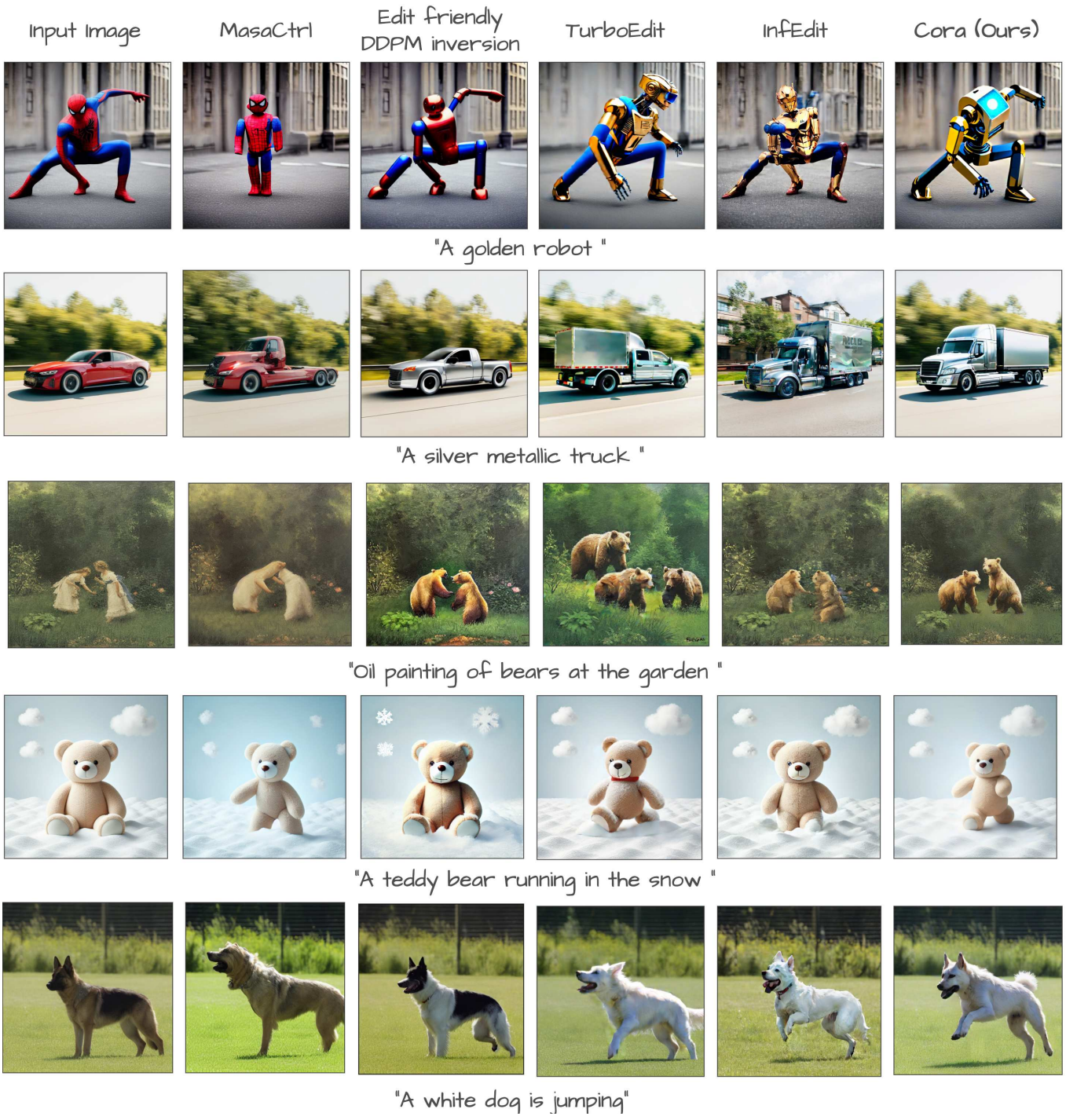


Fig. 9. **Qualitative comparison**. We compare the visual quality of our edits with several few-step and multi-step frameworks. our method strikes the perfect balance between respecting the editing text-prompt and preserving the content of the input image.

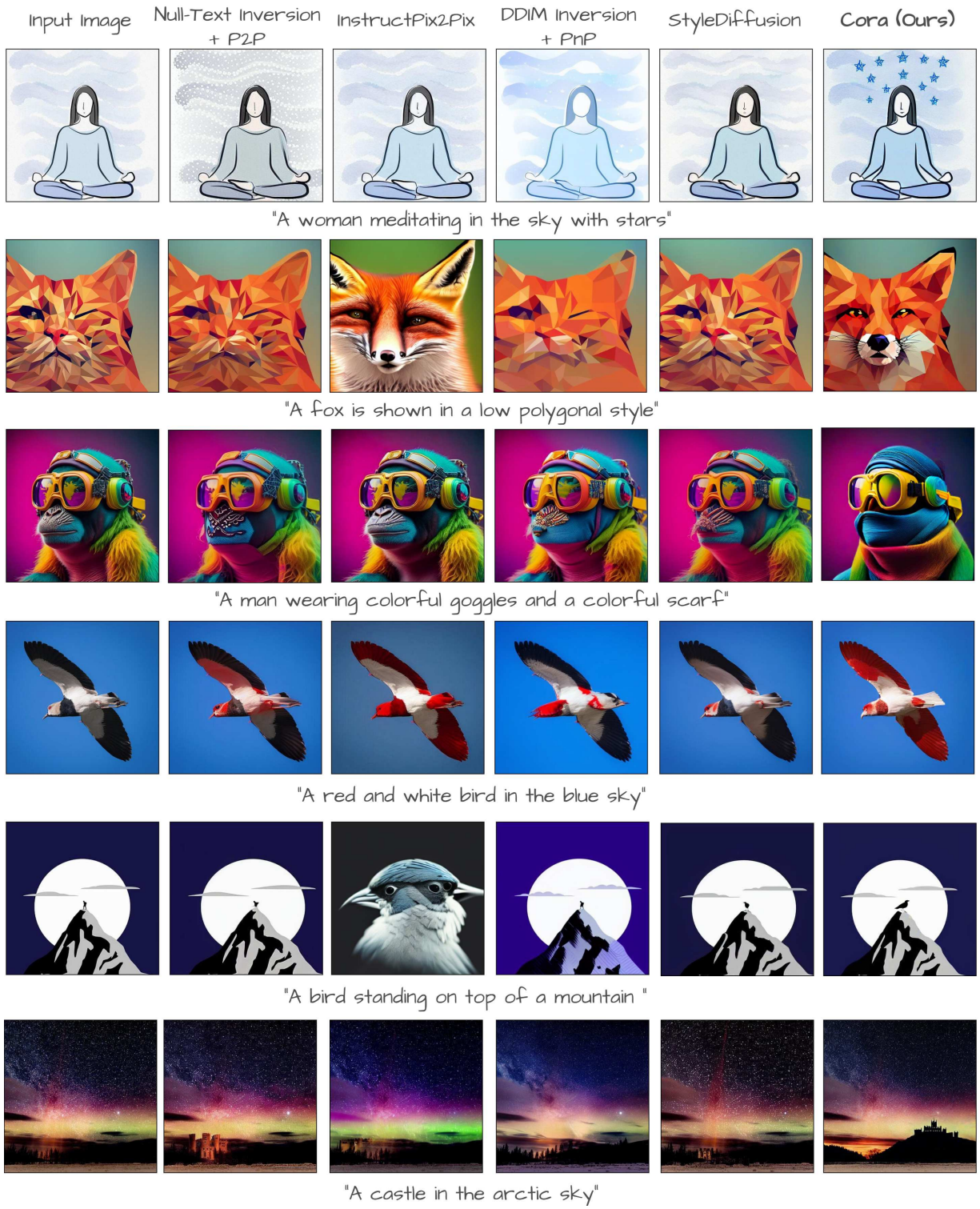


Fig. 10. **Additional Qualitative Comparison.** We compare our method against several editing frameworks, including Null-Text Inversion [Mokady et al. 2023], InstructPix2Pix [Brooks et al. 2023a], Plug-and-Play [Tumanyan et al. 2022], and StyleDiffusion [Li et al. 2023].

A IMPLEMENTATION DETAILS

We use SDXL-Turbo [Sauer et al. 2024] from the Diffusers library [von Platen et al. 2022] as our few-step diffusion model. Both the inversion of the source image and the generation of the target image are performed using *four* denoising steps. Following FlexiEdit [Koo et al. 2024], we observed that high-frequency components in the latent space during the initial timesteps of denoising can hinder flexible editing by imposing rigid constraints, limiting significant deviations from the original image's structure. To facilitate larger and more flexible changes, we apply high-frequency suppression to the latent representation at the first timestep. Furthermore, while our method works without masking, incorporating a mask to specify regions of the image that should remain unchanged can enhance background preservation during editing. Following Avrahami et al. [2022], we use masked latent blending to preserve these regions by blending them from the source image. We applied our latent correction in timesteps 3 and 4 of denoising in which the patch sizes are 5×5 and 3×3 respectively. Moreover, in our content-adaptive interpolation, we found that setting $k = 3$ or $k = 4$ typically produces better results.

B QUANTITATIVE RESULTS

We perform a quantitative comparison of Cora with several few-step and multi-step editing methods. for content preservation, we measure PSNR, LPIPS [Zhang et al. 2018], mean squared error (MSE), and structural similarity [Wang et al. 2004] losses between the backgrounds of the source and target images. For text alignment we measure CLIP similarity [Radford et al. 2021] between the target image with and without subject masking and the editing text-prompt. The results are presented in Tab. 3.

C ADDITIONAL EXPERIMENTS

C.1 Ablation on Structure alignment

In Fig. 12, we show five examples to illustrate how our structural alignment step preserves the original pose and layout. Each row compares results generated *without* alignment (left) to those generated *with* alignment (right). Without alignment, the target images tend to deviate significantly from the source structure. In contrast, applying our one-to-one query matching ensures that the coarse spatial arrangement of the source image is retained, resulting in edited images that faithfully reproduce the same pose while incorporating the desired changes.

C.2 Ablation on latent correction

In Fig. 14, we compare our editing results *with* and *without* applying latent correction. When the correction is skipped, we frequently observe misalignment artifacts and unnatural deformations, especially for edits involving large shape changes. In contrast, incorporating our proposed patchwise correspondence alignment effectively mitigates these artifacts, leading to higher-quality edits that better preserve both the overall structure and fine details of the source image. This highlights the importance of the latent correction mechanism in achieving coherent and visually appealing results.

C.3 Ablation on different attention mixing strategies

In Fig. 15, we provide visual comparison between different attention mixing strategies. These include mutual self-attention [Cao et al. 2023], Concatenating source and target image attentions [Hertz et al. 2023b; Tewel et al. 2024], and our linear and spherical interpolation with and without DIFT alignment. As evident, the aligned spherical interpolation yields the most realistic and natural results.

D GENERALIZATION THROUGH CONTROLLABLE PARAMETERS

When edits rely solely on text instructions, users often find it hard to specify how much of the source's appearance or structure should remain intact. We address this by introducing two user-defined parameters, α and β, and, because our approach only uses a few denoising steps, users can quickly adjust α and β to reach the desired outcome with minimal time overhead. As shown in Figure 16 to Figure 21, tuning α determines how much of the original look is preserved versus newly generated, while β refines how strictly the layout follows the source. These parameters provide clear, fine-grained control, reducing guesswork and enabling a broad range of edits, from slight refinements to large-scale transformations, while preserving stable and predictable outcomes.

E LATENT CORRECTION VIA PATCH-BASED CORRESPONDENCE

E.1 Why the Original Corrections Fail

Noise–inversion of the input image returns a sequence of latent corrections $\{z_t\}_{t=1}^{T}$ such that, if we inject the same z_t's during the backward pass and keep the conditioning text unchanged, the model reconstructs the source image pixel-perfectly. Each z_t therefore assumes that every object will still occupy the exact spatial location it had in the source.

Editing, however, breaks this assumption: for instance, if a dog is asked to jump, the pixels affected by the edit are now displaced, and no longer correspond to the locations for which the original z_t corrections were computed. Reusing those unmodified corrections forces the model to inject noise that is no longer spatially aligned, leading to mismatched textures, silhouette glitches, and the reappearance of content that no longer belongs.

E.2 Solution: Correspondence-aware Latent Corrections

Fig. 11 contrasts the standard TurboEdit pipeline (middle) with our correspondence-aware variant (bottom). After inversion (top) we still obtain the original noise residual z_{t-1}, but before re-injecting it we *realign* it so that it follows the geometry of the edited frame x_{t-1}.

Our patch-based correspondence approach divides DIFT feature maps into overlapping patches, matches each target patch to its most similar source patch, and then reconstructs the aligned target representation by reassembling the matched patches back into their original spatial locations and averaging overlapping regions.

Table 3. Quantitative comparisons among text-based editing baselines. **Bold** indicates the best scoring method, underline indicates the second best, blue indicates the third best.

	PSNR (Background) ↑	LPIPS (Background) ↓	MSE (Background) ↓	SSIM (Background) ↑	CLIP Sim. (Whole) ↑	CLIP Sim. (Edited) ↑
Prompt2prompt [Hertz et al. 2022]	19.9	153.98	188.94	79.58	24.87	22.35
Plug-and-play [Tumanyan et al. 2022]	24.88	80.86	72.53	86.25	**25.13**	22.08
NTI [Mokady et al. 2023]	**29.92**	**36.68**	**25.18**	**91.02**	24.18	21.09
StyleDiffusion [Li et al. 2023]	28.65	45.16	34.53	89.6	24.02	21.28
InstructPix2Pix [Brooks et al. 2023a]	22.47	145.50	241.69	80.38	21.54	19.80
DDPM inversion [Huberman-Spiegelglas et al. 2024]	24.09	110.68	220.50	84.65	23.35	20.40
MasaCtrl [Cao et al. 2023]	24.47	82.81	78.44	87.01	23.99	21.00
TurboEdit [Deutch et al. 2024]	27.95	44.98	37.87	89.82	24.79	22.44
InfEdit [Xu et al. 2024]	25.51	123.76	370.24	80.21	23.34	20.08
Cora (Ours)	28.02	50.57	38.03	89.39	24.91	**22.69**

E.3 Patch Extraction and Matching

Let $D_S \in \mathbb{R}^{C\times H\times W}$ and $D_T \in \mathbb{R}^{C\times H\times W}$ be the feature maps of the source and target images, respectively. We extract overlapping patches of size $k \times k$ from both maps with stride s. This turns each map into a set of flattened patch vectors:

$$\begin{aligned} \text{patches}_S &= \{p_{S,1}, \ldots, p_{S,N_S}\}, \quad p_{S,i} \in \mathbb{R}^{Ck^2}, \\ \text{patches}_T &= \{p_{T,1}, \ldots, p_{T,N_T}\}, \quad p_{T,j} \in \mathbb{R}^{Ck^2}. \end{aligned} \tag{22}$$

Here, N_S and N_T denote the total numbers of patches extracted from the source and target, respectively.

We then compute cosine similarity between every pair $(p_{T,j}, p_{S,i})$:

$$\text{sim}(p_{T,j}, p_{S,i}) = \frac{p_{T,j} \cdot p_{S,i}}{\|p_{T,j}\| \, \|p_{S,i}\|}. \tag{23}$$

For each target patch $p_{T,j}$, we select the index of the best matching source patch:

$$C_{T\to S}(j) = \arg\max_{1\le i\le N_S} \text{sim}(p_{T,j}, p_{S,i}). \tag{24}$$

E.4 Reassembling Matched Patches

After finding the best-match index for each target patch, we replace each target patch $p_{T,j}$ with its matched source patch $p_{S,\mathcal{I}(j)}$, yielding an updated patch set:

$$\hat{p}_{T,j} = p_{S,\, C_{T\to S}(j)}. \tag{25}$$

We then reassemble these patches into a tensor of shape (C, H, W) by placing each $\hat{p}_{T,j}$ at the corresponding spatial location of the j-th patch in the target. Since patches overlap, every pixel can receive contributions from multiple patches. Specifically, we denote by $\tilde{D}_T(\mathbf{u})$ the sum of all matched-patch contributions at spatial location $\mathbf{u}$. Let Ω_j be the set of pixel coordinates covered by the j-th patch. Then:

$$\tilde{D}_T(\mathbf{u}) = \sum_{j:\, \mathbf{u}\in\Omega_j} \hat{p}_{T,j}(\mathbf{u}), \tag{26}$$

$$W(\mathbf{u}) = \sum_{j:\, \mathbf{u}\in\Omega_j} 1, \tag{27}$$

where $\hat{p}_{T,j}(\mathbf{u})$ is the feature value of patch $\hat{p}_{T,j}$ for pixel $\mathbf{u}$. To normalize overlaps, we compute the final aligned map $\hat{D}_T$ by dividing each spatial location by its total overlap:

$$\hat{D}_T(\mathbf{u}) = \frac{\tilde{D}_T(\mathbf{u})}{W(\mathbf{u}) + \varepsilon}, \tag{28}$$

with ε a small constant to prevent division by zero.

F CONTENT-ADAPTIVE INTERPOLATION

When a prompt leads to large deformations or introduces new objects, not every pixel in the edited (target) image should be forced to match a pixel in the original (source) image. Over-enforcing alignment often creates visual artifacts or incorrect texture transfers. To address this, we propose a two-step strategy that checks whether each target patch has a reliable counterpart in the source before blending.

F.1 1. Bidirectional Matching

Let S be the set of source patches and T be the set of target patches. For each source patch $s \in S$, we define its top-k most similar target patches as:

$$\mathcal{K}(s) = \arg\operatorname{top}_k\{\text{sim}(s,t)\}, \tag{29}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity between feature vectors. Similarly, for each $t \in T$, define $\mathcal{K}(t)$. A pair (s,t) is said to be *bidirectionally matched* if $s \in \mathcal{K}(t)$ and $t \in \mathcal{K}(s)$. These are considered strong correspondences, and we blend source and target information using a user-defined weight α.

F.2 2. Weak Matches → Treat as New

Some target patches remain unmatched. For each such target patch $t \in T$, we compute the strength of its best match in the source as:

$$\text{sim}_{\max}(t) = \max_{s\in S} \text{sim}(s,t). \tag{30}$$

Let $U \subset T$ be the set of unmatched patches. We compute the γ-quantile threshold τ_γ over $\text{sim}_{\max}(t)$ values:

$$\tau_\gamma = \text{quantile}_\gamma\left(\{\text{sim}_{\max}(t) \mid t \in U\}\right). \tag{31}$$

If $\text{sim}_{\max}(t) < \tau_\gamma$, we classify t as *new* and set $\alpha = 1$, meaning it is fully guided by the prompt and not the source image.

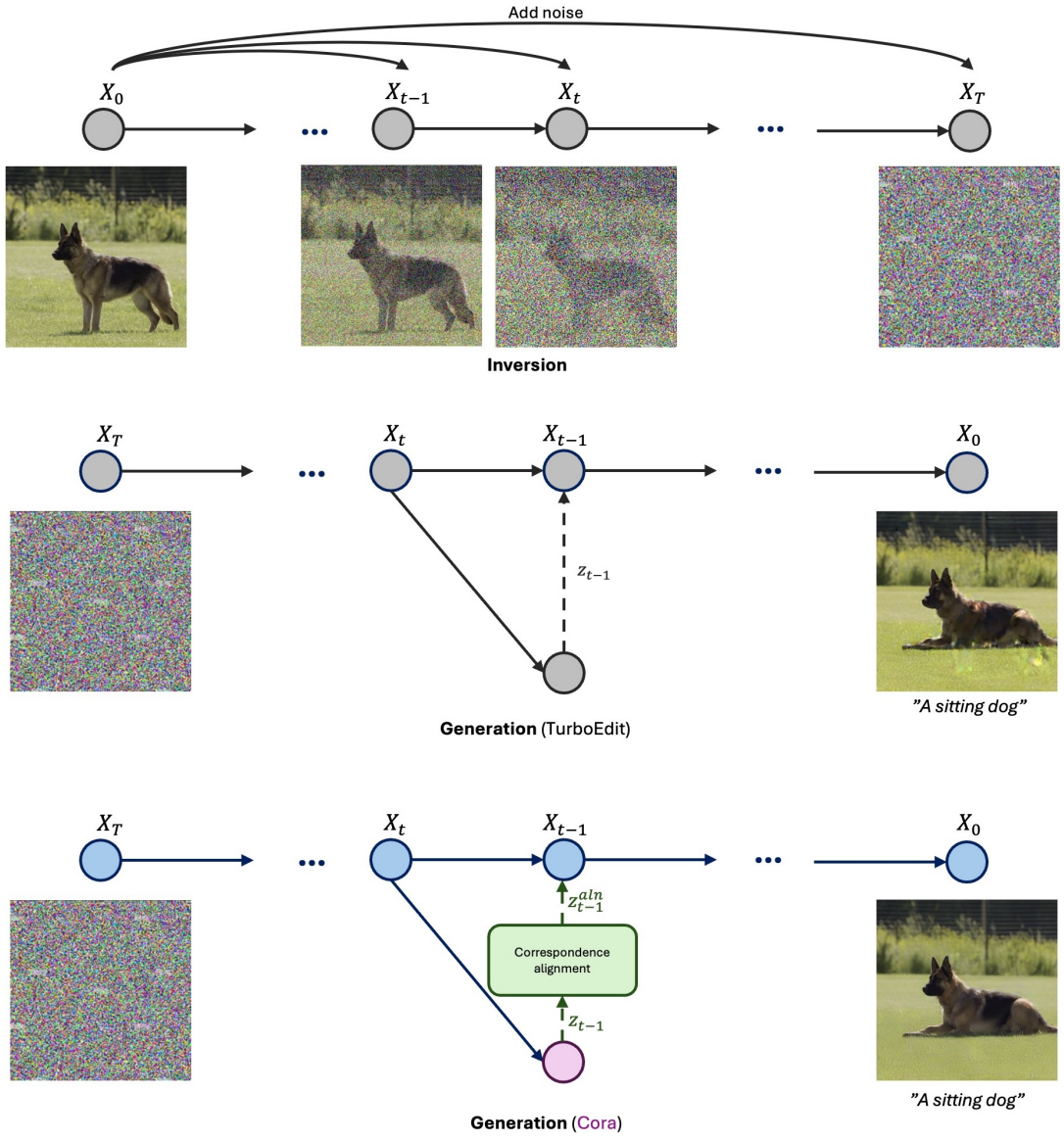


Fig. 11. **Correspondence-aware latent correction.** Top: inversion turns the source image into noise x_T and extracts residuals $\{z_t\}$ that exactly reconstruct the original pose. Middle: TurboEdit re-injects the unaligned residual (correction term) z_{t-1} while editing the pose, causing textures to snap back to old positions. Bottom: our method aligns z_{t-1} via DIFT-based patch correspondences, producing a geometry-aware correction z_{t-1}^{aln} and a clean, artifact-free result.

F.3 Practical Insights

This content-adaptive interpolation balances preservation and generation: we reuse source appearance where reliable correspondences exist, and rely on prompt-driven generation in new regions. This avoids artifacts from over-aligning unrelated content. While the matching process is not perfectly accurate, we found it to be sufficient in most examples. In practice, it significantly reduces artifacts caused by mistakenly interpolating between unmatched patches.

Interestingly, even when only a portion of a new region—such as 20% of the pixels corresponding to a newly generated object like a hat—is correctly identified as "new," the outcome is often satisfactory. Since we avoid interpolation for those pixels, they remain purely prompt-driven. During denoising, the rest of the image tends to adapt around these pixels, effectively completing the structure and appearance of the new content. Nonetheless, improving the accuracy of identifying new regions would further enhance the quality and reliability of the generated results.

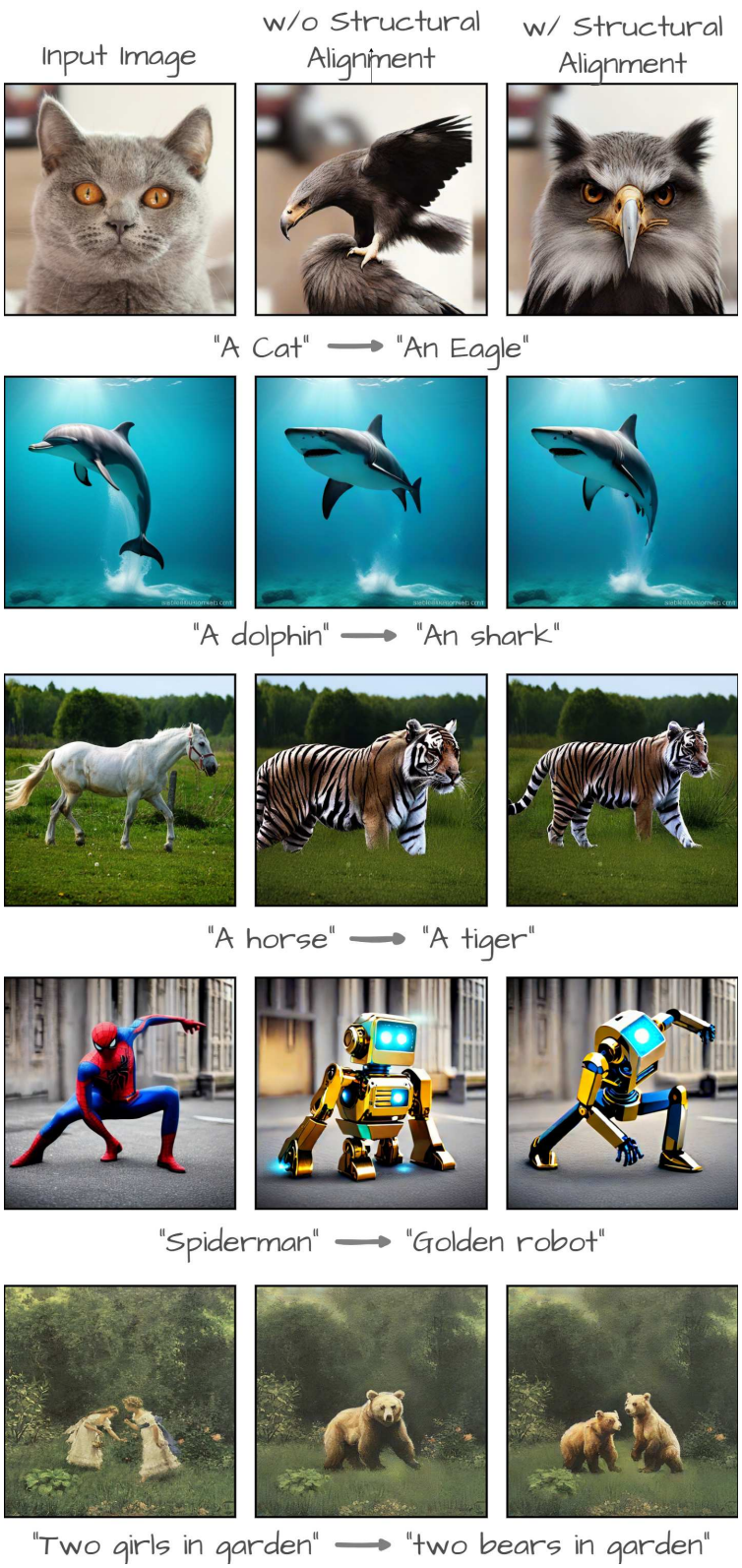


Fig. 12. **Ablation on structure alignment.** By applying our structure alignment, we can preserve the structural layout of the source image.

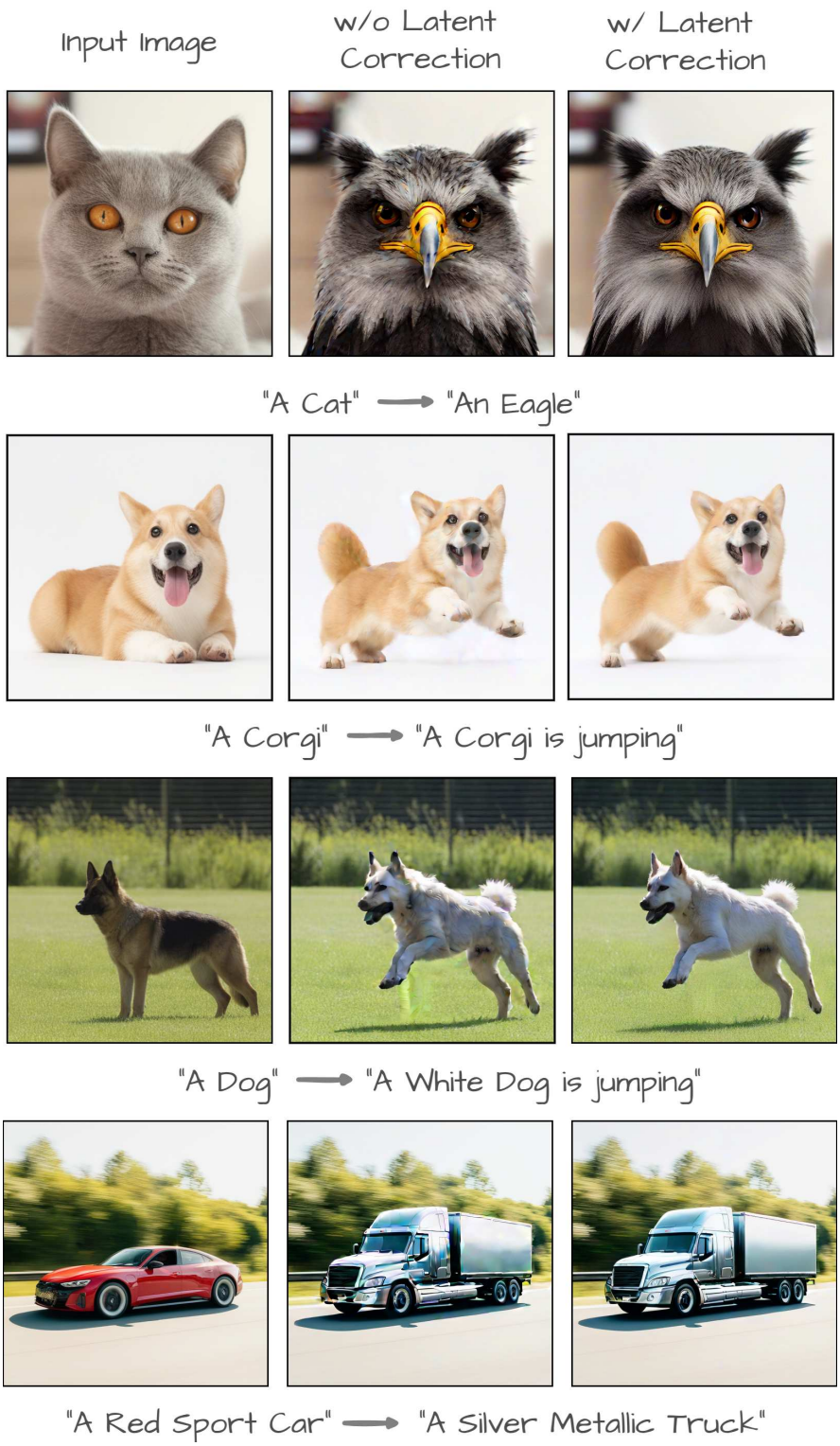


Fig. 13. **Ablation on latent correction.** Without latent correction, multiple misalignment artifacts and unnatural deformations occur. Applying correction produces cleaner and more realistic results.

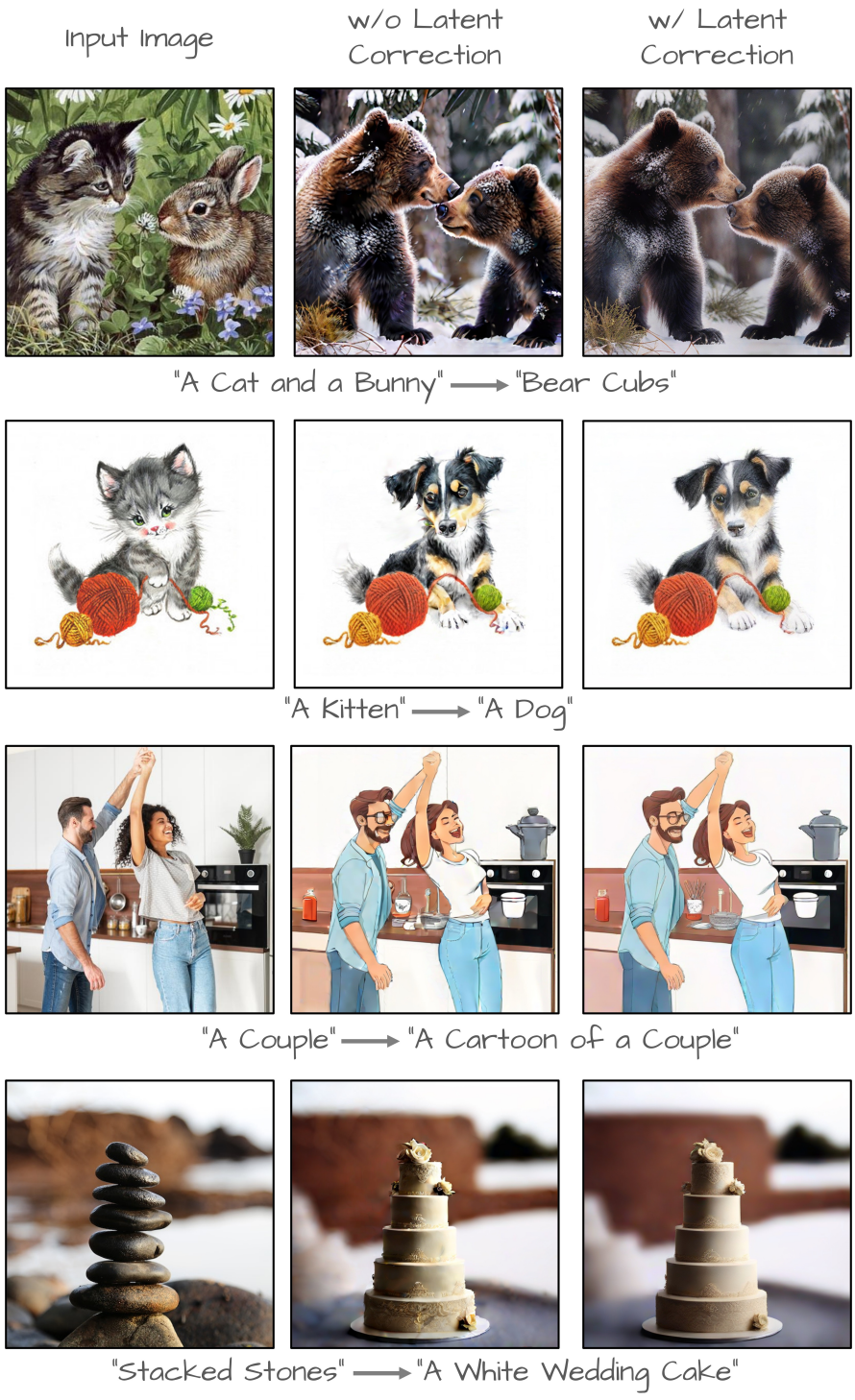


Fig. 14. **Ablation on latent correction.** Without latent correction, multiple misalignment artifacts and unnatural deformations occur. Applying correction produces cleaner and more realistic results.

Fig. 15. **Ablation on attention mixing strategies.** With these visual results, we demonstrate that DIFT-aligned SLERP yields the best results.

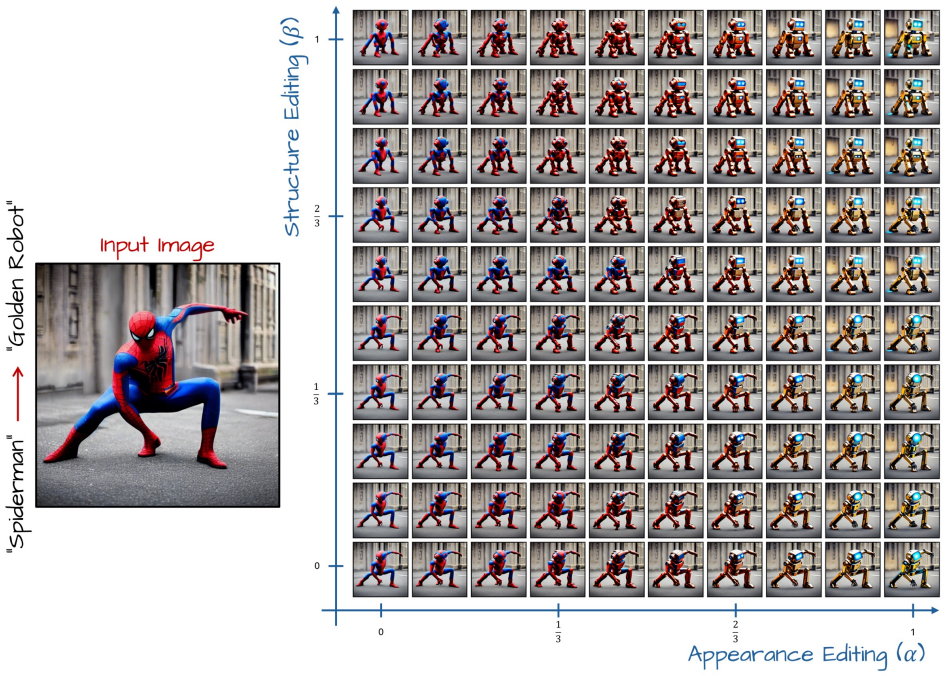


Fig. 16. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.

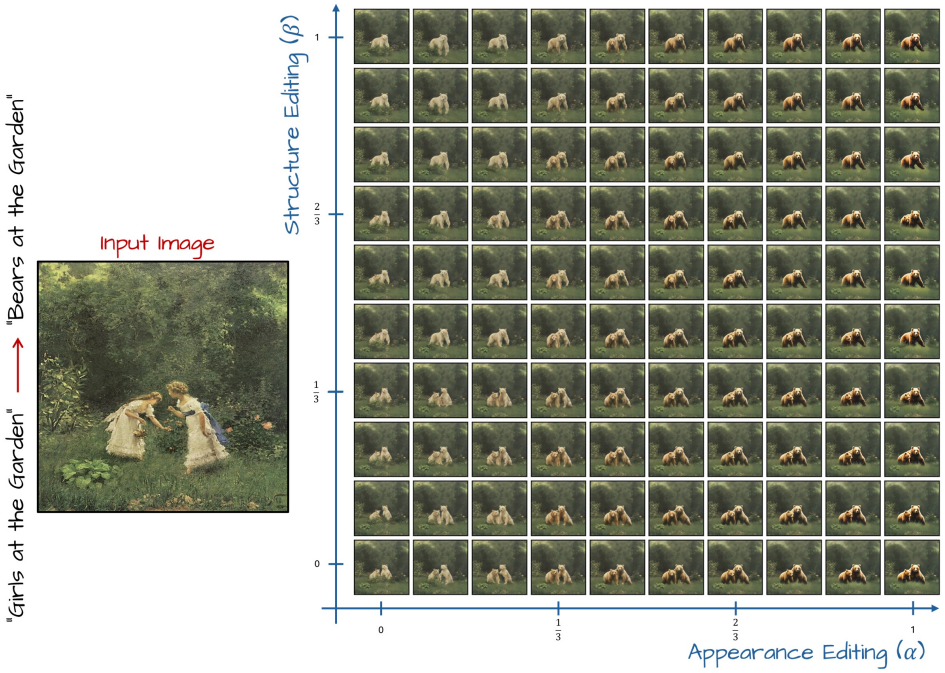


Fig. 17. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.

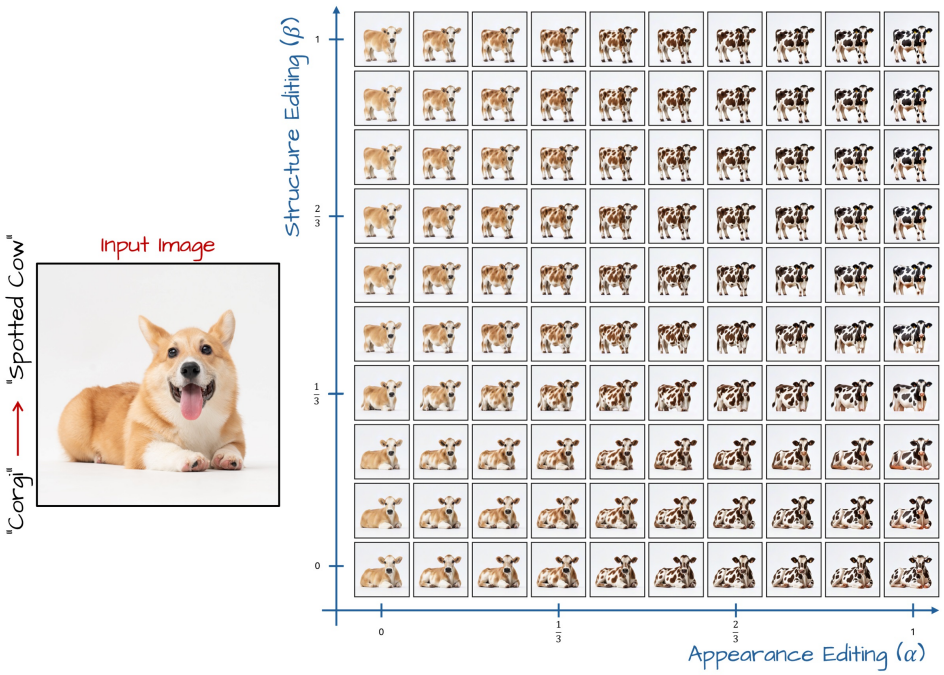


Fig. 18. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.

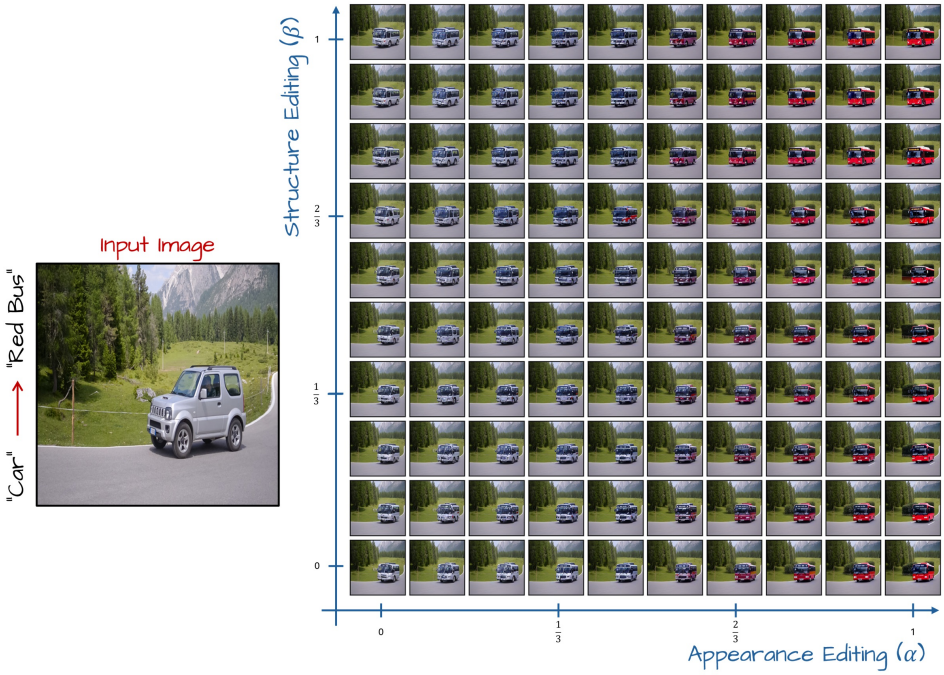


Fig. 19. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.

Fig. 20. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.

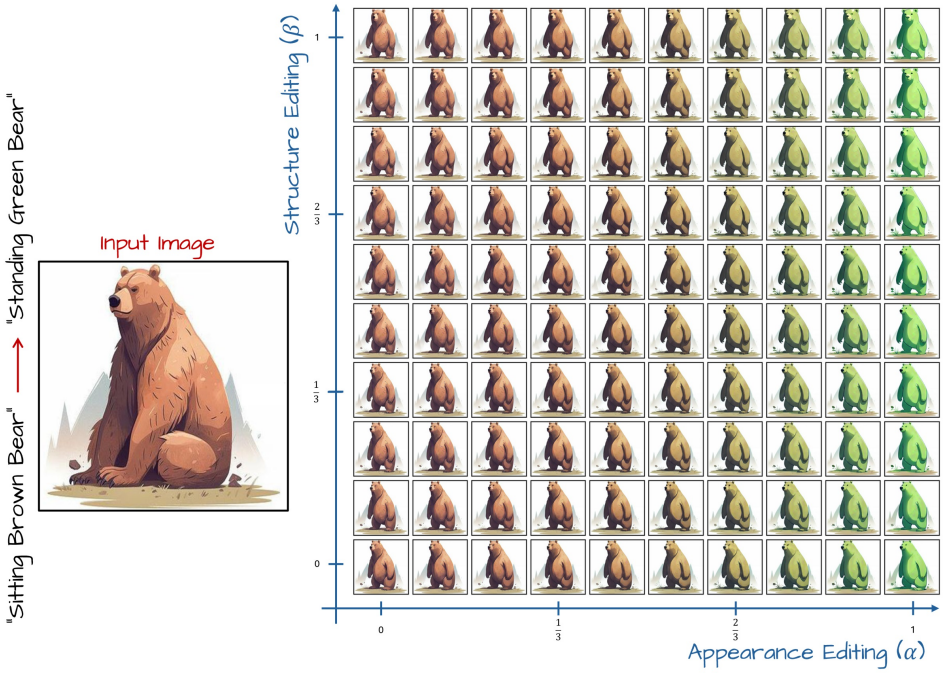


Fig. 21. Additional results showcasing our correspondence-aware attention interpolation and structural alignment. Adjusting α smoothly shifts the appearance from the source to the target, while varying β progressively alters structural elements. The grid shows how appearance and structure can be controlled independently to achieve diverse transformations.