This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Edicho: Consistent Image Editing in the Wild

Qingyan Bai[1,2*] Hao Ouyang[2] Yinghao Xu[3,2] Qiuyu Wang[2]
Ceyuan Yang[4] Ka Leong Cheng[1,2] Yujun Shen[2†] Qifeng Chen[1†]

[1]HKUST [2]Ant Group [3]Stanford University [4]CUHK

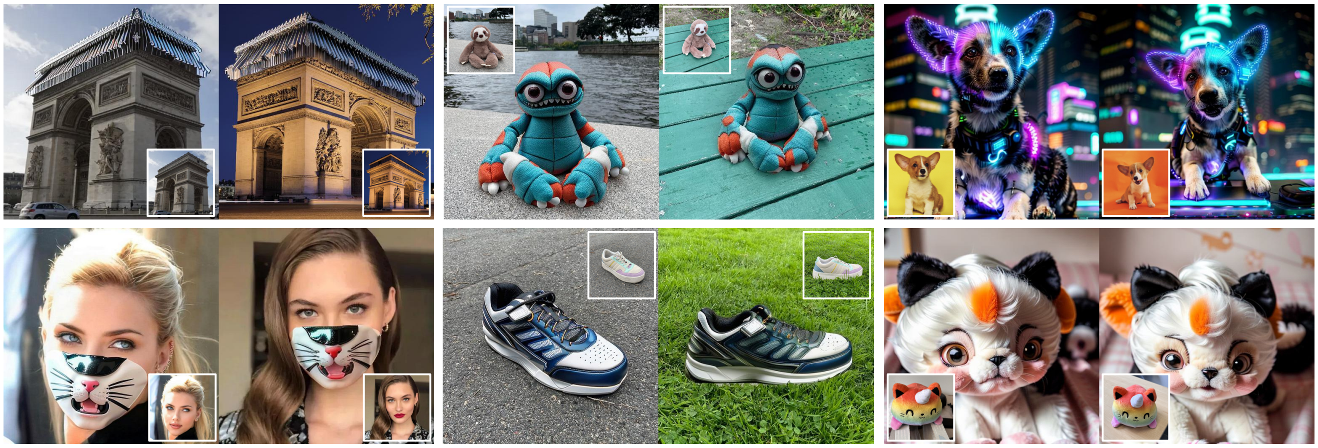

Figure 1. Given two images in the wild, Edicho generates consistent editing versions of them in a zero-shot manner. Our approach achieves precise consistency for editing parts (left), objects (middle), and the entire images (right) by leveraging explicit correspondence.

Abstract

As a verified need, consistent editing across in-the-wild images remains a technical challenge arising from various unmanageable factors, like object poses, lighting conditions, and photography environments. Edicho[1] *steps in with a training-free solution based on diffusion models, featuring a fundamental design principle of using **explicit** image correspondence to direct editing. Specifically, the key components include an attention manipulation module and a carefully refined classifier-free guidance (CFG) denoising strategy, both of which take into account the pre-estimated correspondence. Such an inference-time algorithm enjoys a plug-and-play nature and is compatible to most diffusion-based editing methods, such as ControlNet and BrushNet. Extensive results demonstrate the efficacy of* Edicho *in consistent cross-image editing under diverse settings. Project page can be found here.*

1. Introduction

The ability to consistently edit images across different instances is of paramount importance in the field of computer vision and image processing [1, 10, 65, 67]. Consistent image editing facilitates numerous applications, such as creating coherent visual narratives and maintaining characteristics in marketing materials. As in Fig. 1, sellers or consumers can enhance photos of their favorite products, such as toys or shoes, by applying consistent decorative elements, making each item appear more appealing or personalized. Similarly, during themed events like a masquerade ball or Halloween, families and friends may hope to uniformly style masks or dresses across their photos, ensuring a harmonious visual presentation. Another instance for content creators is consistently making multiple of the photos looks like a graceful elf or an impressive superman. By ensuring the edits applied to one image can be reliably replicated across the other ones, we also enhance the efficiency and quality of tasks ranging from photo editing and retouching to data augmentation for customization [32, 50], and 3D reconstruction [57].

Despite the significance of consistent editing, achieving it across diverse images remains a challenging task. Pre-

∗ This work was done during an internship at Ant Group.
† Corresponding authors.
[1] "Edicho" is an abbreviation of "edit echo", implying that the edit is echoed across images.

vious editing methods [5, 25, 64] often operate on a per-image basis, leading to variations that can disrupt the uniformity required in specific applications. Prior attempts to address this issue have encountered limitations. Learning-based methods [10, 61] that involve editing a single image and propagating the changes to others lack proper regularization and tend to produce inconsistent results. They struggle to acquire high-quality paired training data and fail to enforce the necessary constraints to maintain uniformity. Alternatively, strategies without optimization [1, 6, 18] relies on the implicit attention features to achieve appearance transfer. Yet due to the instability of implicit representations, these approaches struggle to account for the intrinsic variations between images, leading to edits that appear inconsistent or distorted when applied indiscriminately.

Inspired by the property of diffusion models [19, 35, 48, 51] where intermediate features are spatially aligned with the generated image space, we propose a novel, training-free, and plug-and-play method that enhances consistent image editing through explicit correspondence between images. Different from previous training-free methods [1, 6, 18] relying on implicit attention features, we propose to predict the explicit correspondence between the inputs with a robust correspondence extractor before editing. Our approach then leverages the self-attention mechanism within diffusion models to transfer features from a source image to a target image effectively. Specifically, we enhance the self-attention mechanism by warping the query features according to the correspondence between the source and target images. This allows us to borrow relevant attention features from the source image, ensuring that the edits remain coherent across different instances. To achieve finer control over the consistency of the edits, we further modify the classifier-free guidance (CFG) [11] computation by incorporating the pre-computed correspondence. This modification guides the generation process, aligning it more closely with the desired edits while maintaining high image quality. During this design, we empirically observed that directly transferring the source noisy latent to the target image often results in blurred and over-smoothed images. Inspired by the concept of NULL-text Inversion [37], we discovered that fusing features from unconditional embeddings enhances consistency without compromising the image quality.

Moreover, our algorithm is specifically designed to handle in-the-wild images - those captured under diverse and uncontrolled real-world conditions. Benefiting from the correspondence, this capability ensures that our method remains robust against variations in lighting, backgrounds, perspectives, and occlusions commonly found in natural settings. By effectively processing the wild images, the versatility of our method allows for additional numerous practical applications. For instance, in customized generation, our method enables the generation of more consistent image sets by editing, which is valuable for learning customized models for novel concepts and creating personalized content. Additionally, we can apply new textures consistently across different views of an object and acquire the corresponding 3D reconstructions of the edits, benefiting from the editing consistency.

In summary, we introduce explicit correspondence into the denoising process of diffusion models in order to achieve consistent image editing. We enhance the self-attention mechanism and modify classifier-free guidance to incorporate correspondence information, improving edit consistency without degrading image quality. We also further demonstrate that fusing features from unconditional embeddings enhances consistency, inspired by Null-text Inversion techniques. The final method, due to its training-free and plug-and-play nature, is able to function across various models and diverse tasks, enabling both global and local edits. We validate the effectiveness of the proposed method through extensive experiments, showing superior performance in both quantitative metrics and qualitative assessments.

2. Related Works

Generative models for image editing. Recently, diffusion models have shown unprecedented power in various generative tasks [2–4, 8, 11, 13–15, 20, 21, 28–32, 34, 40–42, 48–51, 53, 56, 58, 59, 62]. To unleash its potential in editing, PnP [54] proposes to borrow convolutional and attention features from the input image during generation to achieve manipulation. While MasaCtrl [6] and Cross-Imgae-Attention [1] modify self-attention modules for editing, by combining the target queries and source keys and values. Prompt2Prompt [17] focuses on the cross-attention layers in text-to-image models and proposes manipulating the textual embedding. Serving as the foundation of these editing methods, image inversion is also widely studied by researchers [16, 26, 36, 37, 51]. Different from the aforementioned training-free editing methods, InstructPix2Pix [5], ControlNet [64, 66], T2I-Adapter [38], Composer [23], and BrushNet [25] learn editing models conditioned on the input images and instructions, which are based on or fine-tuned from the pre-trained latent diffusion models for better quality and training stability. Another branch of works [9, 10, 61] aims at exemplar-based editing, where a pre-trained diffusion model is finetuned to function conditioned on the exemplar image as well as the masked source image. [39] achieves pose transfer among the image batch by manipulating the StyleGAN latent codes following the exemplar image. Unlike the works discussed above, we focus on the task of consistent editing for images in the wild, and propose an explicit correspondence-guided solution.

Correspondence from neural networks. The concept of correspondence is widely applicable and essential in various real-world scenarios [65, 67], where understanding relationships between data points is crucial. Neural networks have been broadly employed to find correspondences for image, video, and 3D scenes through supervised learning [24, 24, 33, 45, 57]. DIFT [52] proposes to extract semantic correspondence among in-the-wild images by directly matching the features from the pre-trained diffusion models. SD-DINO [63] further ensembles features from diffusion models and DINO [7] for correspondence matching. Once correspondences are established, they can be utilized in various applications. For instance, in object tracking, networks can maintain correspondences across video frames to follow objects through occlusions and transformations [12, 27, 41, 44, 60]. Our work leverages these principles by integrating correspondence into the diffusion model framework, enabling precise and consistent multi-image editing without additional training.

3. Method

In this work, we focus on the task of consistent image editing, where multiple images are manipulated altogether to achieve consistent and unified looks. To achieve this, we first extract explicit semantic correspondence among the image pairs by existing visual understanding methods such as [33, 52, 57, 63]. Then we seek help from the pre-trained editing models [25, 64] built upon Stable Diffusion [48] to achieve editing, and guide their denoising process with these pre-computed explicit correspondence to ensure consistency. In this section, we first review some preliminary concepts of diffusion models, which is followed by a subsection discussing correspondence prediction and analysis. Then we introduce the correspondence-guided denoising process that includes two levels - the level of attention features and the level of noisy latents in CFG. Note that these designs on feature manipulations are only applied to a range of denoising steps and layers, in order to preserve the strong generative prior from the pre-trained models.

3.1. Preliminaries

Diffusion models are probabilistic generative models trained through a process of progressively adding and then removing noise. The forwarding process adds noise to the images as follows:

$$\boldsymbol{x}_t = \sqrt{\alpha_t} \cdot \boldsymbol{x}_0 + \sqrt{1-\alpha_t} \cdot \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$ and α_t indicates the noise schedule. And a neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is trained to predict the adding noise $\boldsymbol{z}$ during the denoising backward process and finally achieves sampling from Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \mathbf{I})$. In the formulation of latent diffusion models (LDMs) [48], a pair of pre-trained variational encoder $\mathcal{E}$ and decoder $\mathcal{D}$ serve perceptual compression and enable denoising from the noisy latents $\mathbf{z}$ in this latent space.

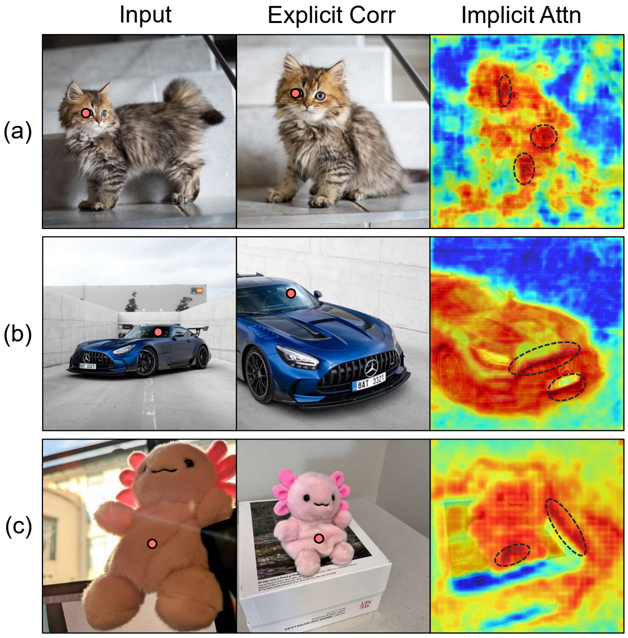

Figure 2. **Visualization** of the explicit correspondence and implicit attention maps for the images in the wild. The implicit features obtained from attention calculation are less accurate and unstable. Regions with the highest attention weights are outlined with dashed circles.

Classifier-free guidance (CFG) [11] represents a groundbreaking method designed to improve both the quality and diversity of images generated by diffusion models, without the need for additional classifiers. By incorporating a mixing coefficient, CFG effectively combines the conditional and unconditional predictions from the denoising model. The unconditional prediction is typically derived by setting the condition to a null or default value.

Reference networks for editing. Recent advances in editing techniques [25, 64] have introduced a novel approach by implementing an auxiliary reference network over pre-trained large diffusion models, while maintaining the fixed architecture of the pre-trained backbone. This network-topology-preserving design ensures a clear separation between control signals and the pre-trained generative prior, enabling more precise and flexible editing capabilities.

3.2. Correspondence Prediction and Analysis

Correspondence prediction. To achieve consistent editing for images I_i and I_j, we first extract explicit correspondence and study the comparison between it and the implicit features. The correspondence from the input images are acquired with a pre-trained correspondence extractor such as [52, 57]:

$$\mathcal{C}_{i,j} = \phi(I_i, I_j), \tag{2}$$

where ϕ and $\mathcal{C}$ indicate the extractor and correspondence.

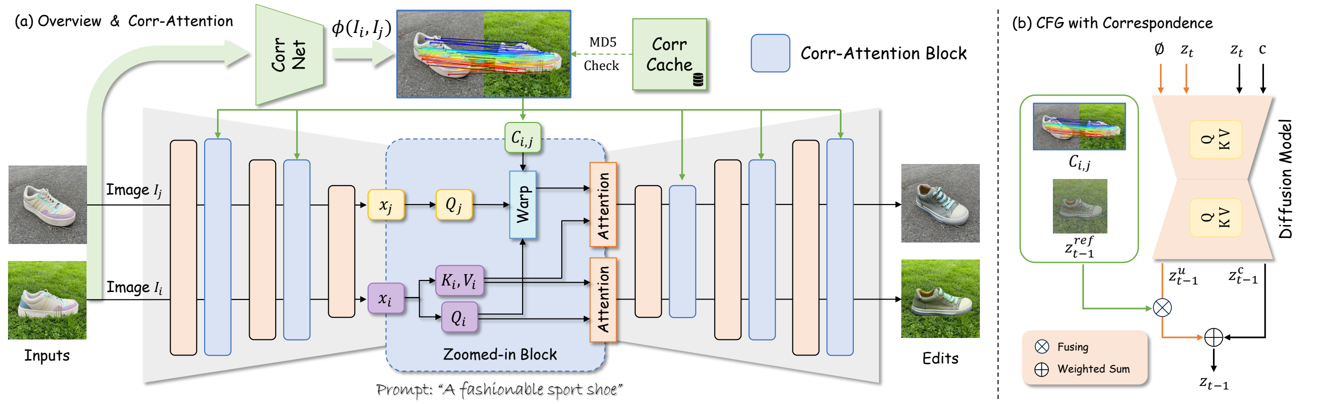

Figure 3. **Framework** of Edicho. To achieve consistent editing, we first predict the explicit correspondence with extractors for the input images. The pre-computed correspondence is injected into the pre-trained diffusion models and guide the denoising in the two levels of (a) attention features and (b) noisy latents in CFG.

Analysis. We further conduct comparisons between the explicit correspondence and implicit features. Explicit extractors predict correspondence from the input images with a single-pass forwarding and apply this prediction to all the target network layers and denoising steps during editing in our design. While implicit extractors predict correspondence by calculating the similarities between attention queries and keys for each of the layer and denoising step, and subsequently apply it in this layer and step, as in the previous training-free editing methods. Specifically, We employed DIFT [52] for explicit prediction. As for the implicit approach, we follow Cross-Image-Attention [1] to compute the attention similarity by first querying the attention keys of the matching image with $Q_i \cdot K_j^T$, where i and j indicate the image indices. We then select a point on the source image and compute the attention map based on the aforementioned attention similarity. In Fig. 2, we present the correspondence prediction results using explicit and implicit methods. The regions with the highest attention weights in Fig. 2 are outlined with dashed circles, which suggests the implicit methods would query unreasonable regions that would cause inconsistent textures, while the correspondences obtained by explicit prediction are notably more accurate than those obtained by implicit methods. As a result, for the editing methods merely based on implicit attentions [1, 6, 18], the inaccurate matching leads to borrowing inaccurate features in performing cross-image attention, which hinders the editing consistency. This further boost our motivation of introducing more robust explicit correspondence to guide the denoising process.

Efficiency optimization. To further optimize the efficiency of obtaining explicit correspondence, we implement a strategy to avoid redundant computations, particularly when the same images or image groups are processed multiple times. We achieve this by encoding each image group using an MD5 [47] hash function, creating a unique identifier for each. After storing the identifier (key) and correspondence (value) in a minor database, the input image group would first retrieve it before editing for acceleration.

3.3. Attention Manipulation with Correspondence

Recall that the intermediate feature $\mathbf{x_i}$ in self-attention blocks are firstly projected to queries $Q_i = f_Q(\mathbf{x_i})$, keys $K_i = f_K(\mathbf{x_i})$, and values $V_i = f_V(\mathbf{x_i})$ with the learned projection matrix f_Q, f_K, and f_V. Then the attention features F could be computed by autonomously computing and assessing the relevance of these feature representations following [55]. Inspired by the comparisons between explicit and implicit representations, we propose to guide self-attention with explicit correspondence to achieve consistent editing, which is termed as Corr-Attention. As in Fig. 3, for an image pair (I_i, I_j) among the inputs, we borrow features from the query matrix Q_i to Q_j to form a new query Q_{edit} based on the explicit correspondence:

$$Q_{edit} = \mathit{Warp}(Q_i, Q_j, \mathcal{C}_{i,j}), \tag{3}$$

Where the *Warp* function indicates the process of borrowing features by warping corresponding tokens to the source based on the corresponding location denoted by correspondence. Considering (1) tokens of Q_{edit} are borrowed from Q_i and (2) to further improve consistency, we querying K_i, V_i instead of K_j, V_j during the editing of I_j:

$$F_j = \mathrm{softmax}\left(\frac{Q_{edit} \cdot K_i^T}{\sqrt{d_k}}\right) \cdot V_i, \tag{4}$$

where d_k indicates the dimension of Q and K, and F_j represents the attention outputs of I_j. By transferring attention features from the source, we effectively achieve editing consistency during the denoising process.

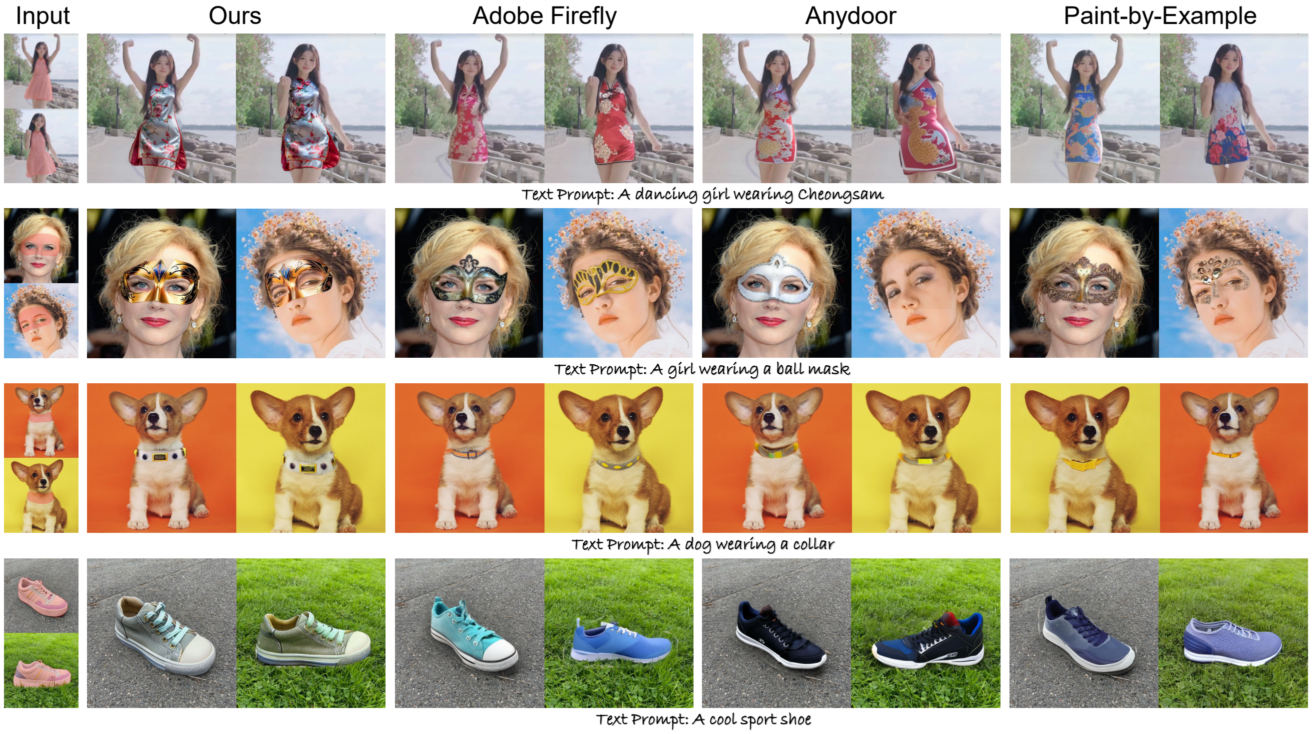
Input
Ours
Adobe Firefly
Anydoor
Paint-by-Example
Text Prompt: A dancing girl wearing Cheongsam
Text Prompt: A girl wearing a ball mask
Text Prompt: A dog wearing a collar
Text Prompt: A cool sport shoe

Figure 4. Qualitative comparisons on local editing with Adobe Firefly (AF) [46], Anydoor (AD) [10], and Paint-by-Example (PBE) [61]. The inpainted areas of the inputs are highlighted in red.

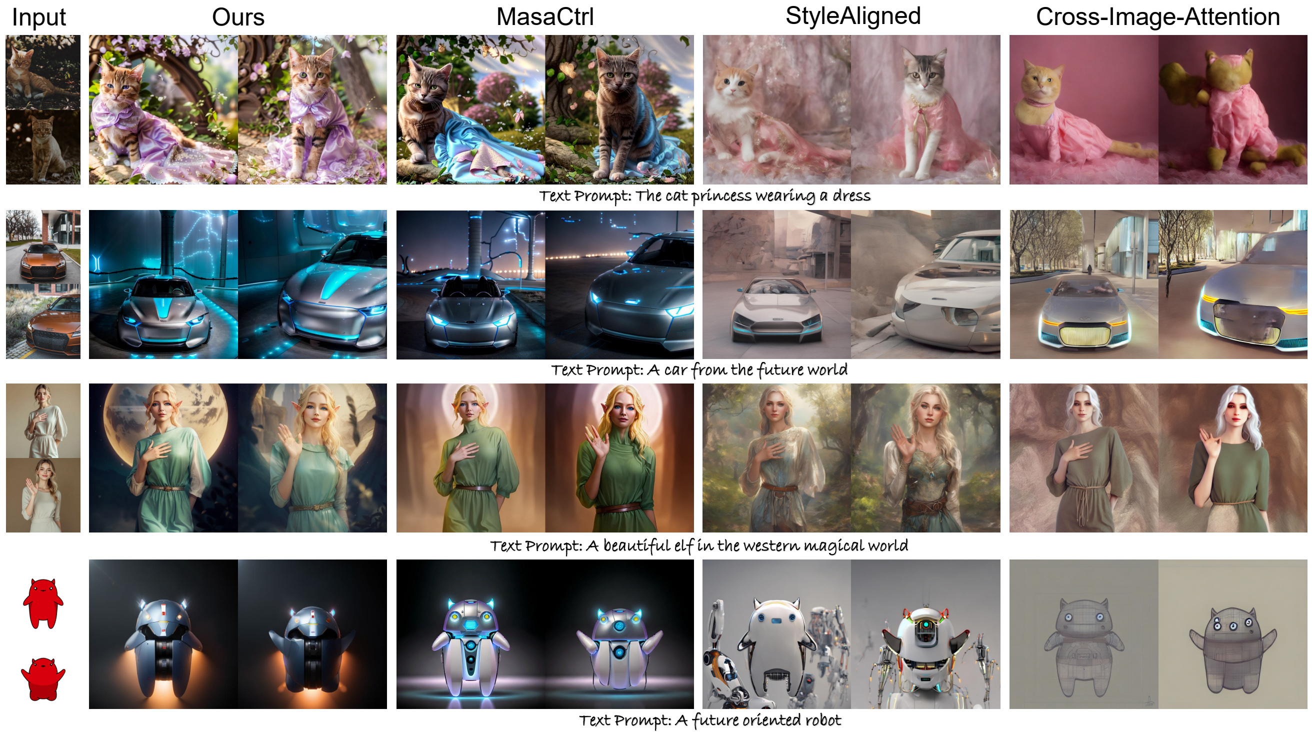
Input
Ours
MasaCtrl
StyleAligned
Cross-Image-Attention
Text Prompt: The cat princess wearing a dress
Text Prompt: A car from the future world
Text Prompt: A beautiful elf in the western magical world
Text Prompt: A future oriented robot

Figure 5. Qualitative comparisons on global editing with MasaCtrl (MC) [6], StyleAligned (SA) [18], and Cross-Image-Attention (CIA) [1].

3.4. Classifer-free Guidance with Correspondence

In order to retain a finer consistency over the edited images, we take a further step from the attention feature control and focus on the noisy latents in Classifier-free Guidance (CFG). Specifically, we extend the traditional CFG framework to facilitate synchronized editing of multiple images by leveraging explicit correspondences and propose Corr-CFG. NULL-text Inversion [37] demonstrates that optimizing unconditional word embeddings can achieve precise image inversion and semantic editing. Inspired by this approach, our primary objective is to preserve the integrity of the pre-trained model's powerful generative priors during the consistent editing process. To achieve this, we propose manipulating only the unconditional branch of $\mathbf{z_j}$ within the Classifier-Free Guidance (CFG) framework under the guidance of correspondence, as in the right panel of Fig. 3. Recall that in CFG, the denoising process is split into two branches: conditional and unconditional, and the noise is estimated using a neural network ϵ_θ:

$$\mathbf{z}_{t-1}^{c} = \mathbf{z}_{t}^{c} - \epsilon_\theta\left(\mathbf{z}_t, c\right), \tag{5}$$

$$\mathbf{z}_{t-1}^{u} = \mathbf{z}_{t}^{u} - \epsilon_\theta\left(\mathbf{z}_t, \varnothing\right), \tag{6}$$

where c represents the condition (text prompt) and $\varnothing$ indicates the null text. Specifically, we modify the unconditional noise prediction of $\mathbf{z_j}$ and incorporate information from noise prediction of $\mathbf{z_i}$ into it during the denoising process, which ensures coherent edits:

$$\epsilon_\theta^{u}\left(\mathbf{z}_j\right) = \mathcal{T}\left(\epsilon_\theta\left(\mathbf{z}_i, \varnothing\right), \epsilon_\theta\left(\mathbf{z}_j, \varnothing\right), \mathcal{C}_{i,j}\right), \tag{7}$$

where $\mathcal{T}$ represents a fusing function that aligns the unconditional noises and t indicates the time-step:

$$\mathcal{T}(\mathbf{p}, \mathbf{q}, \mathcal{C}) = (1-\lambda)\cdot\mathbf{q} + \lambda\cdot Inj(\mathbf{p}, \mathbf{q}, \mathcal{C}, \gamma). \tag{8}$$

λ and $\gamma \in (0, 1]$ here are adjustable parameters. The function Inj indicates a process for choosing the source latents from $\mathbf{p}$ and injecting them into the corresponding locations of target latents $\mathbf{q}$, based on the correspondence $\mathcal{C}$. The factor γ here is introduced to modulate the proportion of injected target latents, serving as a balancing mechanism between preserving the generative prior and maintaining editing consistency throughout the process. At last, we apply the guidance and fuse the conditional and unconditional predictions as in the prior paradigm [11]:

$$\epsilon_\theta^{\text{guided}}\left(\mathbf{z}_t, c\right) = \epsilon_\theta^{u}\left(\mathbf{z}_t\right) + s\cdot\left(\epsilon_\theta\left(\mathbf{z}_t, c\right) - \epsilon_\theta^{u}(\mathbf{z}_t)\right), \tag{9}$$

where s indicates the guidance scale. The final latents generated as such are at last sent to the VAE decoder [48] to be decoded into images.

4. Experiment

4.1. Experimental Setup

Settings. We use Stable Diffusion [48] as the base model and adopt BrushNet [25] and ControlNet [64] as the reference networks for editing. We adopt the DDIM [51] scheduler and perform denoising for 50 steps. By default, the proposed correspondence-guided denosing strategy is applied from 4^{th} to 40^{th} steps and from the eighth attention layer to ensure consistency as well as preserve the strong generative prior. Note the optimal choice of these may vary when different base models are utilized. The testing samples are partially acquired from the internet, while others of them are from the dataset of DreamBooth [50] and Custom Diffusion [32].

Evaluation metrics. We follow Custom Diffusion [32] and adopt the prevalent multi-modal model CLIP [43] to evaluate various methods in terms of text alignment (TA) and editing consistency (EC). Specifically, on the one hand, feature similarity of the target prompt and model output are computed to judge the textual alignment. On the other hand, feature similarity of the edited images are adopted to evaluate the editing consistency. User studies (US) are also incorporated to further evaluate the practical applicability and user satisfaction.

Baselines. We include both local and global editing tasks, as well as numerous previous image editing methods for comprehensive comparisons. Specifically, for the task of local editing, we include prior works of Adobe Firefly [46], Anydoor [10], and Paint-by-Example [61] for comparison. Among these aforementioned methods, Firefly is a state-of-the-art commercial inpainting tool developed by Adobe, which could repaint local regions of the input image following the given textual prompts. In order to achieve the task of consistent editing, the images among the set would be inpainted with the same detailed prompts. Both of Anydoor and Paint-by-example are Latent Diffusion Models (LDMs) supporting repaint target region with the given reference image. Thus we sent an inpainted image to these models as the reference, expecting consistent editing results. While for global editing, we compare our approach with MasaCtrl [6], StyleAlign [18], and Cross-Image-Attention [1]. The aforementioned methods achieve editing by manipulating and fusing attention features from various sources. Different from our method, they rely on implicit attention weights to ensure consistency among the editing outputs.

4.2. Evaluation

Qualitative results. We present a qualitative evaluation of the consistency editing methods, focusing on both local editing (image inpainting) and global editing (image translation). The comparisons for local editing in Fig. 4 include results from our method, Adobe Firefly (AF), Anydoor

Table 1. Quantitative results respectively on local and global editing. We follow Custom Diffusion [32] to evaluate various methods on text alignment (TA) and editing consistency (EC).

Method	TA ↑	EC ↑	Method	TA ↑	EC ↑
AF [46]	0.3082	0.8569	MC [6]	0.3140	0.9258
AD [10]	0.2981	0.8320	SA [18]	0.3021	0.9099
PBE [61]	0.2969	0.8683	CIA [1]	0.2914	0.8912
Ours	**0.3176**	**0.8931**	Ours	**0.3228**	**0.9355**

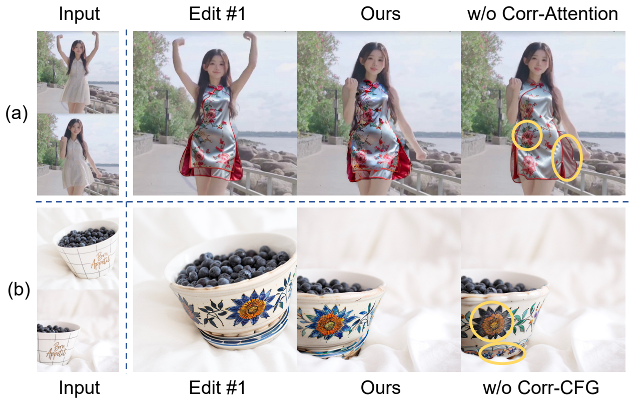

Figure 6. Ablation studies on the (a) correspondence-guided Attention Manipulation (Corr-Attention) and (b) correspondence-guided CFG (Corr-CFG).

(AD), and Paint-by-Example (PBE). The results demonstrate that our approach consistently maintains the integrity of the input images across different modifications including the cloth textures, mask and collar appearance, and even the eyelet amount of shoes thanks to the introduction of explicit correspondence. The baselines of global editing mainly include the ones predicting merely by implicit attentions - MasaCtrl (MC), StyleAligned (SA), and Cross-Image-Attention (CIA). As in Fig. 5, our method also achieves superior consistency and thematic adherence among the edits such as the dress of the cat. While the implicit alternatives such as MasaCtrl fails in the car roof, the high neckline of the elf, and the hole number of the robot.

Quantitative results. We conducted a comprehensive quantitative evaluation of our proposed method against several state-of-the-art image editing techniques, focusing on metrics of text alignment (TA) and editing consistency (EC) mentioned in Sec. 4.1. As in Tab. 1, for local editing, our method attained the best scores in both TA and EC for local editing tasks, demonstrating a significant improvement over the competing methods. For global editing tasks, our method continued to outperform the other counterparts, reaching a TA score of 0.3228 and EC score of 0.9355. Overall, these results along with the user studies in *Supplementary Material* clearly demonstrate the effectiveness of our method in achieving high text alignment and editing consistency across both local and global editing scenarios.

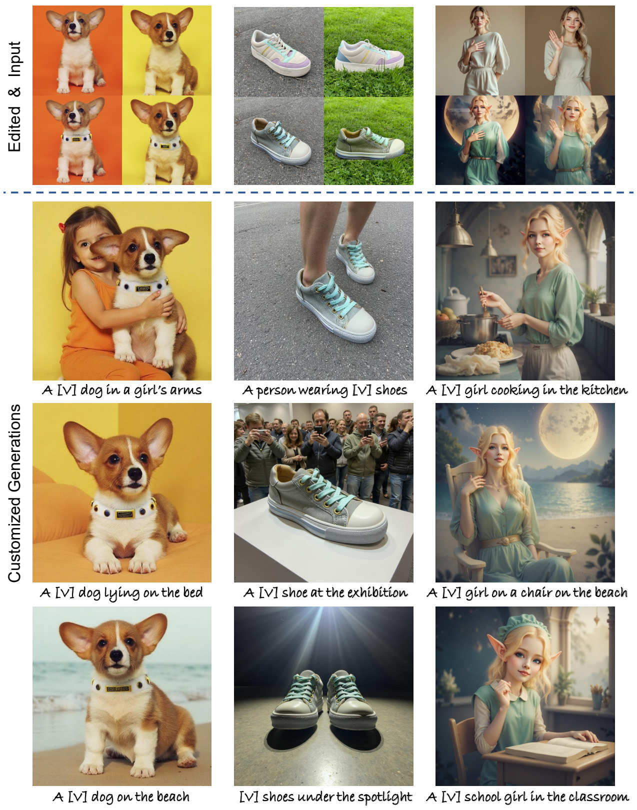
Edited & Input
Customized Generations
A [V] dog in a girl's arms
A person wearing [V] shoes
A [V] girl cooking in the kitchen
A [V] dog lying on the bed
A [V] shoe at the exhibition
A [V] girl on a chair on the beach
A [V] dog on the beach
[V] shoes under the spotlight
A [V] school girl in the classroom

Figure 7. With outputs from our consistent editing method (upper) and the customization [50] techniques, customized generation (lower) could be achieved by injecting the edited concepts into the generative model.

4.3. Ablation Studies

In order to validate the effectiveness of the proposed correspondence-guided attention manipulation (Corr-Attention) and correspondence-guided CFG (Corr-CFG) introduced in Secs. 3.3 and 3.4, we conduct ablation studies by respectively disabling each of them and testing on the task of consistent editing. When the proposed correspondence-guided attention manipulation (Corr-Attention) is disabled, the diffusion model relies on implicit attention to maintain consistency just like previous methods [1, 6, 18]. As demonstrated in Fig. 6a, the generative model then would yield flowers of wrong amounts and at improper locations. The number of flowers and inconsistent textures speak for the effectiveness of introducing explicit correspondence to attention manipulation. Recall that correspondence-guided CFG (Corr-CFG) is designed for finer consistency control functioning in the latent space of LDMs, which is validated in Fig. 6b where Corr-CFG achieves in generating more consistent textures for the flower on the bowl and stripes at the bottom of the bowl. Additional ablation studies on the attention manipu-

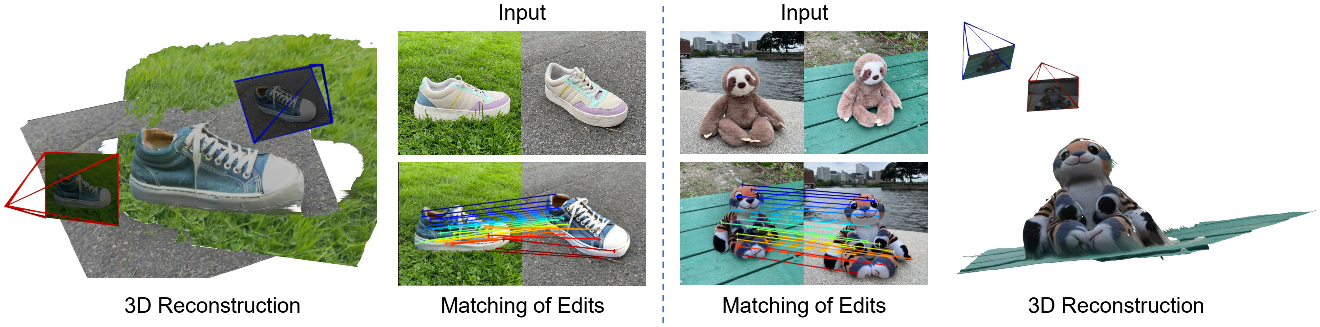

Figure 8. We adopt the neural regressor Dust3R [57] for 3D reconstruction based on the edits by matching the 2D points in a 3D space.

lation and CFG could be found in *Supplementary Material*.

4.4. Additional Applications

Customization based on the consistent edits. To further demonstrate the practical utility of the proposed method, we present an application example that integrates DreamBooth [50] and Low-Rank Adaptation (LoRA) [22] techniques for customized image generation based on the multiple edited images. Leveraging the edited outputs from our method, we employ DreamBooth to fine-tune the diffusion-based generative model for 500 steps for concept injection. We also integrate LoRA techniques into the learning to further enhance the efficiency of this process by introducing low-rank matrices as adaptation parameters. As in Fig. 7, the fine-tuned generative model could yield desirable images corresponding to the edits after concept injection. This demonstrated synergy between consistent editing and parameter-efficient customization opens new possibilities for content creation.

3D reconstruction based on the consistent edits. Furthermore, consistent editing could also benefit 3D reconstruction of the edits. We achieve 3D reconstruction with a neural regressor [57] which could predict accurate 3D scene representations from the consistent image pairs. Taking the edited images as inputs, the learned neural regressor could predict the 3D point-based models and 2D matchings without additional inputs such as camera parameters. The reconstruction and matching results are presented in Fig. 8, both of which also suggest the editing consistency of the proposed method. The regressor respectively obtained 11,515 and 13,800 pairs of matching points for the two groups of edits, and a portion is visualized for clear understanding. Notably, recovering accurate 2D matches and 3D topology directly from edited images suggests that our approach maintains photometric and geometric consistency.

Additional results. With the proposed correspondence-guided attention and CFG manipulations, additional diverse results of multi-image editing by the proposed method are provided in Fig. 9, showcasing the editing consistency and the diversity of the method's generation.

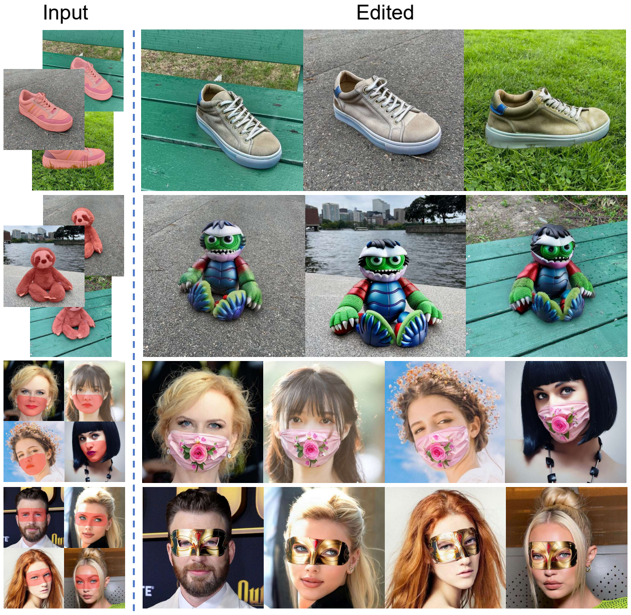

Figure 9. Diverse results of consistent editing for multiple images.

5. Conclusion

We introduce Edicho, a novel training-free method for consistent image editing across various images by leveraging explicit correspondence. Our approach optimizes the self-attention mechanism and the classifier-free guidance computation by integrating correspondence information into the denoising process to ensure consistency. The plug-and-play nature of our method allows for seamless integration into various models and its applicability across a wide range of tasks. Experimental results demonstrate that our method outperforms existing approaches both quantitatively and qualitatively, showcasing its effectiveness in handling diverse and in-the-wild images. For limitations, sometimes the generated textures would be inconsistent due to the correspondence misalignment, which could be expected to be improved with better correspondence extractors. And inheriting from the pre-trained editing models, sometimes distorted textures would be generated.

Acknowledgment

This work was supported by Ant Group Research Intern Program and the National Key R&D Program of China under grant number 2022ZD0161501.

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 1, 2, 4, 5, 6, 7

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 2023. 2

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

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. MasaCtrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Int. Conf. Comput. Vis.*, 2023. 2, 4, 5, 6, 7

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 3

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2

[9] Xi Chen, Yutong Feng, Mengting Chen, Yiyang Wang, Shilong Zhang, Yu Liu, Yujun Shen, and Hengshuang Zhao. Zero-shot image editing with reference imitation. *arXiv preprint arXiv:2406.07547*, 2024. 2

[10] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 2, 5, 6, 7

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Adv. Neural Inform. Process. Syst.*, 2021. 2, 3, 6

[12] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Int. Conf. Comput. Vis.*, 2023. 3

[13] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Int. Conf. Comput. Vis.*, 2023. 2

[14] Peng Gao, Le Zhuo, Dongyang Liu, Ruoyi Du, Xu Luo, Longtian Qiu, Yuhang Zhang, Rongjie Huang, Shijie Geng, Renrui Zhang, et al. Lumina-t2x: Scalable flow-based large diffusion transformer for flexible resolution generation. In *Int. Conf. Learn. Represent.*, 2025.

[15] Xin Gu, Ming Li, Libo Zhang, Fan Chen, Longyin Wen, Tiejian Luo, and Sijie Zhu. Multi-reward as condition for instruction-based image editing. *arXiv preprint arXiv:2411.04713*, 2024. 2

[16] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *IEEE Winter Conf. Appl. Comput. Vis.*, 2024. 2

[17] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Int. Conf. Learn. Represent.*, 2023. 2

[18] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2, 4, 5, 6, 7

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, 2020. 2

[20] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Adv. Neural Inform. Process. Syst.*, 2022. 2

[21] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Int. Conf. Comput. Vis.*, 2023. 2

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Int. Conf. Learn. Represent.*, 2022. 8

[23] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023. 2

[24] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *Int. Conf. Comput. Vis.*, 2021. 3

[25] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *Eur. Conf. Comput. Vis.*, 2024. 2, 3, 6

[26] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *Int. Conf. Learn. Represent.*, 2024. 2

[27] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Eur. Conf. Comput. Vis.*, 2024. 3

[28] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Adv. Neural Inform. Process. Syst.*, 2022. 2

[29] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023.

[30] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Int. Conf. Comput. Vis.*, 2023.

[31] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 2, 6, 7

[33] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Eur. Conf. Comput. Vis.*, 2024. 3

[34] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. In *Int. Conf. Learn. Represent.*, 2025. 2

[35] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2022. 2

[36] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 2

[37] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 6

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Assoc. Adv. Artif. Intell.*, 2024. 2

[39] Thao Nguyen, Utkarsh Ojha, Yuheng Li, Haotian Liu, and Yong Jae Lee. Edit one for all: Interactive batch image editing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[40] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Int. Conf. Mach. Learn.*, 2021. 2

[41] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. CoDeF: Content deformation fields for temporally consistent video processing. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3

[42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Int. Conf. Learn. Represent.*, 2023. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, 2021. 6

[44] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv:2307.01197*, 2023. 3

[45] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[46] Adobe reseachers. Adobe firefly: Free generative ai for creatives. `https://firefly.adobe.com/generate/inpaint`, 2023. 5, 6, 7

[47] Ronald Rivest. The md5 message-digest algorithm, 1992. 4

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 3, 6

[49] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *Int. Conf. Learn. Represent.*, 2025.

[50] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 1, 6, 7, 8

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Int. Conf. Learn. Represent.*, 2021. 2, 6

[52] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Adv. Neural Inform. Process. Syst.*, 2023. 3, 4

[53] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Trans. Graph.*, 2024. 2

[54] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 4

[56] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 2

[57] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 3, 8

[58] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[59] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In *Adv. Neural Inform. Process. Syst.*, 2024. 2

[60] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3

[61] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2, 5, 6, 7

[62] Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. Textctrl: Diffusion-based scene text editing with prior guidance control. In *Adv. Neural Inform. Process. Syst.*, 2024. 2

[63] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *Adv. Neural Inform. Process. Syst.*, 2024. 3

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, 2023. 2, 3, 6

[65] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 3

[66] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Adv. Neural Inform. Process. Syst.*, 2024. 2

[67] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 3