UIP2P: Unsupervised Instruction-based Image Editing via Edit Reversibility Constraint

Enis Simsar[1] Alessio Tonioni[3] Yongqin Xian[3] Thomas Hofmann[1] Federico Tombari[2,3]
[1]ETH Zürich - DALAB [2]Technical University of Munich [3]Google Switzerland
https://uip2p.github.io

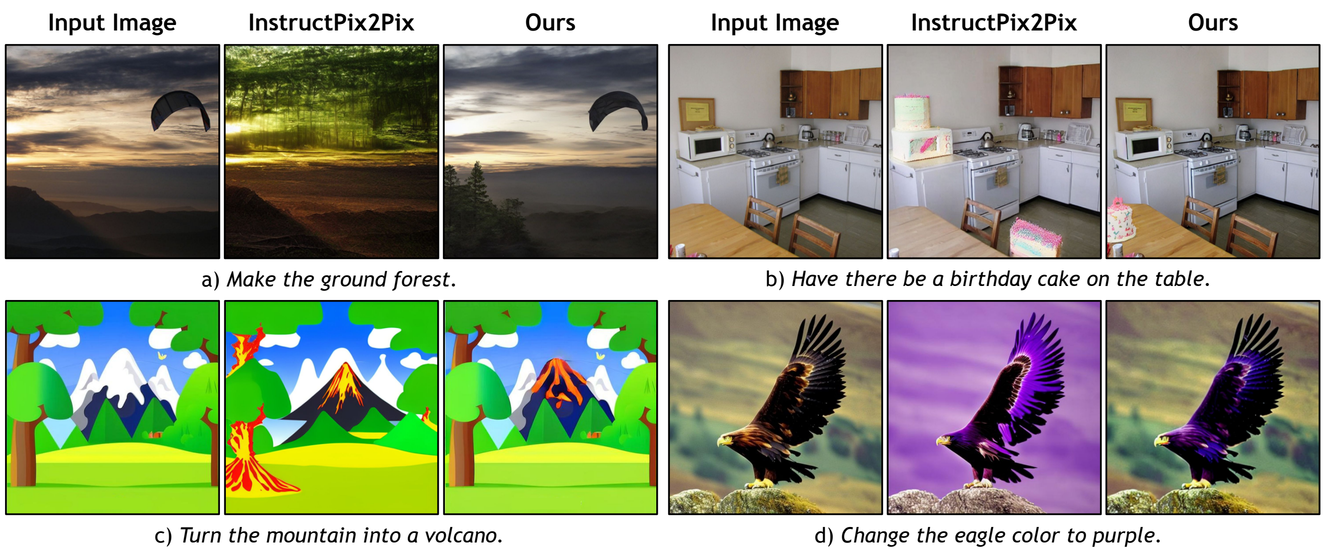


Figure 1. **Unsupervised InstructPix2Pix.** Our approach applies more precise and coherent edits while better preserving scene structure. UIP2P surpasses the supervised alternative, IP2P, trained on a synthetic dataset, demonstrating superior performance on both real (a, b) and synthetic (c, d) images.

Abstract

We propose an unsupervised instruction-based image editing approach that removes the need for ground-truth edited images during training. Existing methods rely on supervised learning with triplets of input images, ground-truth edited images, and edit instructions. These triplets are typically generated either by existing editing methods—introducing biases—or through human annotations, which are costly and limit generalization. Our approach addresses these challenges by introducing a novel editing mechanism called Edit Reversibility Constraint (ERC), which applies forward and reverse edits in one training step and enforces alignment in image, text, and attention spaces. This allows us to bypass the need for ground-truth edited images and unlock training for the first time on datasets comprising either real image-caption pairs or image-caption-instruction triplets. We empirically show that our approach performs better across a broader range of edits with high-fidelity and precision. By eliminating the need for pre-existing datasets of triplets, reducing biases associated with current methods, and proposing ERC, our work represents a significant advancement in unblocking scaling of instruction-based image editing.

1. Introduction

Diffusion models (DMs) have recently made significant advancements in generating high-quality and diverse images, driven largely by breakthroughs in text-to-image generation [17, 33, 34, 36]. This progress has enabled various techniques for tasks such as personalized image generation [11, 35, 46], context-aware inpainting [24, 27, 50], and image editing based on textual prompts [2, 8, 15, 21, 25]. Beyond prompt-based editing, DMs have demonstrated strong capabilities in instruction-based editing, where edits are guided by explicit natural language commands, such as instructions [5, 10, 18, 19, 51, 52].

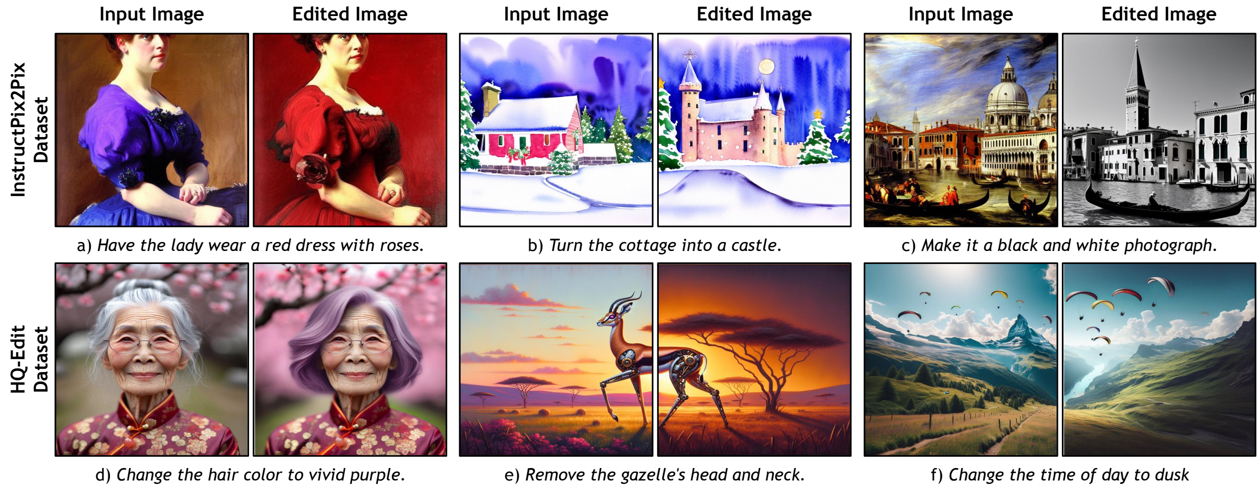


Figure 2. **Supervision Biases in InstructPix2Pix and HQ-Edit datasets.** Each example shows an input image and its corresponding ground-truth edited image for the given edit instruction. InstructPix2Pix employs Prompt-to-Prompt, while HQ-Edit uses DALL-E 3 and GPT-4V. *(a & d) Attribute-entangled edits:* Modifying a specific feature, such as clothing or hair color, unintentionally alters surrounding textures or elements. *(b & e) Scene-entangled edits:* Transforming objects, like turning a cottage into a castle or removing an element, affects unintended parts of the scene. *(c & f) Global changes:* Edits like converting an image to black and white or changing the time of day introduce widespread scene modifications, often compromising visual preservation.

However, existing methods predominantly rely on supervised learning with ground-truth edited images, *i.e.*, requiring datasets of triplets containing input images, edited images, and edit instructions [5, 10, 19, 51, 52]. These datasets are often generated using existing editing methods, such as Prompt-to-Prompt [15], or more recently, advanced diffusion and multimodal models like DALL-E 3 [3] and GPT-4V [1], as in the HQ-Edit dataset [19]. While automated approaches enable scalable dataset creation, they introduce biases (see Fig. 2), including (a & d) attribute-entangled edits that unintentionally alter surrounding elements, (b & e) scene-entangled edits, and (c & f) global changes that significantly modify the entire scene. Conversely, while valuable, human-annotated data [51] is impractical for large-scale training due to the high cost and effort involved. Reliance on automated or human-generated ground-truth edits limits the diversity of achievable modifications, constraining the development of models capable of accurately understanding and executing user instructions.

We present UIP2P, an unsupervised training for instruction-based image editing that eliminates the reliance on triplet datasets, whether generated or human-annotated, by introducing the Edit Reversibility Constraint (ERC)—a constraint enforced while training through forward and reverse edits. Instead of relying on explicitly supervised edited images, we ensure that modifications remain precise and follow the given instructions by leveraging CLIP's ability to align text and images in a shared embedding space [32]. Additionally, we explicitly enforce alignment between forward and reverse edits in both the image and attention spaces, enabling UIP2P to accurately interpret and localize user instructions while ensuring edits remain coherent and reflect the intended changes. ERC allows UIP2P to preserve the integrity of the original content while making precise adjustments, further enhancing the reliability of the edits. Some previous methods have enforced cycle consistency within predefined domain transformations [22, 41, 47, 49, 55]. In contrast, our constraint applies to any arbitrary pair of forward and reverse edits within a single training step, leveraging the expressive power of modern diffusion models to ensure flexibility across diverse editing tasks. By eliminating the dependence on pre-existing datasets, our approach enables training on large-scale real-image datasets—a capability previously constrained by the limitations of existing methods and the high cost of human labeling. As a result, UIP2P significantly expands the scope and scalability of instruction-based image editing compared to prior approaches. Our key contributions are as follows:

- We propose UIP2P, an unsupervised training approach for instruction-based image editing that eliminates the need for ground-truth edited images. By doing so, UIP2P offers a more scalable solution while reducing the biases inherent in alternative supervised methods.
- We introduce the Edit Reversibility Constraint (ERC), a novel mechanism that ensures edits remain stable when applying both forward and reverse edits, preserving structural integrity in the image, text and attention space. This enables precise, high-fidelity modifications that accurately reflect user instructions.

- Our approach demonstrates scalability and versatility across various real-image datasets, allowing a broad range of edits without relying on pre-existing datasets, significantly expanding the capabilities of instruction-based image editing.

2. Related Work

CLIP-Based Image Manipulation. StyleCLIP [29] combines StyleGAN and CLIP for text-driven image manipulation, requiring optimization for each specific edit. Similarly, StyleGAN-NADA [12] enables zero-shot domain adaptation by using CLIP guidance to modify generative models. While these approaches allow for flexible edits, they often rely on domain-specific models or optimization processes for each new task. These works illustrate the potential of CLIP's powerful semantic alignment for image manipulation, which motivates the use of CLIP in other generative frameworks, such as diffusion models.

Text-based Image Editing with Diffusion Models. One common approach in image editing is to use pre-trained diffusion models by first inverting the input image into the latent space and then applying edits through text prompts [8, 15, 20, 25, 26, 28, 44, 45, 48]. For example, DirectInversion [20] edits the image after inversion using Prompt-to-Prompt [15], but the inversion step can lead to losing essential details from the original image. Additionally, methods like DiffusionCLIP [22], CycleDiffusion [47], CycleNet [49], and DualDiffusion [41] explore *domain-to-domain translation* as a way to propose image editing, by using textual prompt or without any conditions. However, their focus on translating between two fixed domains makes it difficult to handle complex edits, such as the insertion or deletion of objects. In contrast, our approach is designed for general-purpose instruction-based image editing without being restricted to domain translation, enabling greater flexibility in handling a broader range of modifications.

Instruction-based Image Editing with Diffusion Models. Another line of methods for image editing involves training models on datasets containing triplets of input image, edit instruction, and edited image [5, 19, 51, 52]. These methods, since they directly take the input image as a condition, do not require an inversion step. InstructDiffusion [13] builds on InstructPix2Pix by handling a wider range of vision tasks but has difficulty with more advanced reasoning. MGIE [10] improves on this by using large multimodal language models to generate more precise instructions. SmartEdit [18] goes a step further by introducing a Bidirectional Interaction Module that better connects the image and text features, improving its performance in challenging editing scenarios. UltraEdit [53] scales instruction-based editing to larger datasets and models, demonstrating improved performance on fine-grained editing tasks.

A major challenge in instruction-based image editing is the need for large-scale, high-quality triplet datasets. InstructPix2Pix [5] partially addresses this by generating extensive datasets using GPT-3 [6] and Prompt-to-Prompt [15]. However, while this mitigates data scarcity, it introduces issues like model biases from Prompt-to-Prompt. MagicBrush [51] tackles the quality aspect with human-annotated datasets, but this approach is small-scale, limiting its practicality for broader use.

Our method leverages CLIP's semantic space for aligning images and text, providing a more robust solution. Moreover, Edit Reversibility Constraint (ERC) tackles both dataset limitations and biases by enforcing coherence between forward and reverse edits. Our approach enhances scalability and precision for complex instructions and eliminates the dependency on triplet datasets, making it applicable to any image-caption dataset of real images. Moreover, as ERC modifies only the training objectives of InstructPix2Pix, it integrates seamlessly with any model extension.

3. Background

Our method builds upon InstructPix2Pix (IP2P) [5], which uses Stable Diffusion's U-Net architecture [34] for instruction-based image editing. IP2P conditions the model on both input images and text instructions using classifier-free guidance [16]. Cross-attention mechanisms compute attention maps based on text conditioning, enabling localized edits by focusing on relevant image regions.

IP2P trains on triplets of input images, edit instructions, and edited images generated using Prompt-to-Prompt [15]. However, this reliance on synthetic data introduces limitations: (1) poor generalization to real-world images, and (2) inherited biases from Prompt-to-Prompt, as shown in Fig. 2.

4. Method

We introduce UIP2P, an unsupervised diffusion-based framework for instruction-based image editing, enforcing the Edit Reversibility Constraint (ERC) to ensure that applied edits can be undone using corresponding reverse instructions. Unlike prior methods reliant on synthetic triplets or domain-specific transformations, UIP2P operates on real-world image-captions datasets without needing edit pairs. Our approach integrates text-image alignment, attention-based localization, reconstruction fidelity, and efficient noise prediction, enabling precise and generalizable edits. Additionally, we leverage LLM-generated reverse instructions, eliminating the need for requiring supervision while significantly improving scalability. In the following sections, we describe our approach in detail, including the key components of our framework (Sec. 4.1), the loss functions that enforce edit reversibility, and the training data generation procedure (Sec. 4.2).

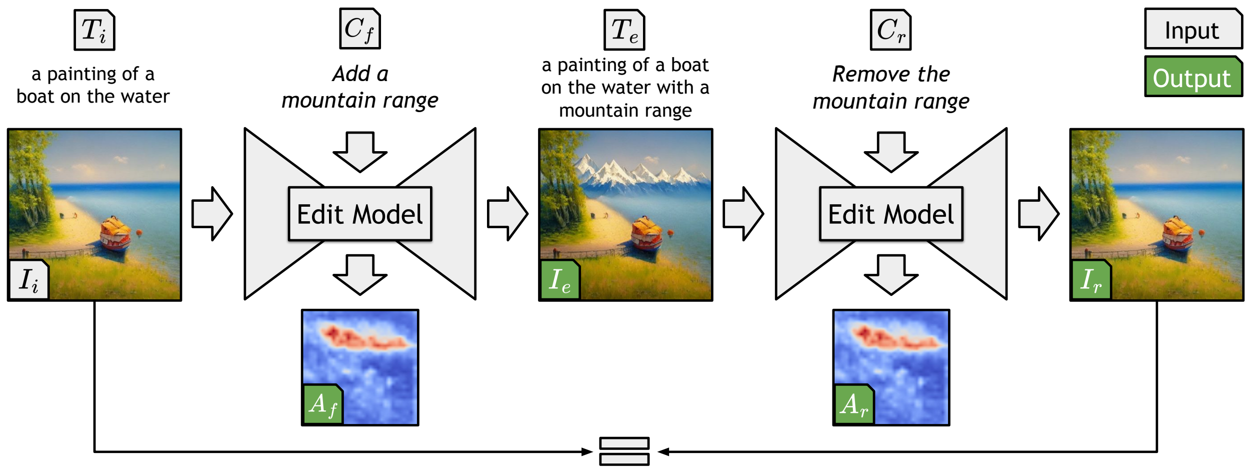


Figure 3. **Overview of the UIP2P training framework.** The model learns instruction-based image editing by applying forward and reverse instructions. Starting with an input image and a forward instruction, shared *Edit Model* generates an edited image. A reverse instruction is then applied to reconstruct the original image, enforcing Edit Reversibility Constraint (ERC).

4.1. Framework

4.1.1. UIP2P

To enforce the Edit Reversibility Constraint (ERC), we define each training step as a sequence of a forward edit followed by its corresponding reverse edit, both applied using the same *Edit Model*. This ensures that modifications can be undone accurately, embedding edit constraints directly within the training process, see Fig. 3 for an overview. A key advantage of our approach is that training is performed *in a single denoising step*, significantly reducing computational cost compared to existing methods that require multiple refinement passes per edit. To enforce the semantic correctness of ERC, we introduce three core components:

1. **Text and Image Direction Alignment**: We leverage CLIP embeddings [31] to align the semantic relationship between textual instructions and image modifications without the need of ground-truth edited images. By operating within CLIP's embedding space, our model ensures that changes in the input and edited images correspond to changes in their respective captions. This alignment is critical for enforcing the Edit Reversibility Constraint (ERC), ensuring that edits accurately follow instructions while preserving the input image's structure.
2. **Attention Map Alignment**: To ensure stable and well-localized edits, we enforce alignment between cross-attention maps generated during both forward and reverse edits. This ensures the model focuses on the same regions of the image when applying and undoing modifications. Attention Map Alignment also regularizes the training objective, improving spatial precision in edits.
3. **Reconstruction Alignment**: A key aspect of ERC is the ability to reconstruct the original input after applying the reverse instruction. This ensures that the model reliably undoes its edits. We achieve this by minimizing both pixel-wise and semantic discrepancies between the reconstructed image and the original input, reinforcing the accuracy of the applied and reversed edits.

During training, we sample a *noise level* t from a predefined schedule and apply it to the input image, I_i, to obtained the noisy version. Based on the *forward instruction*, C_f, *Edit Model* then predicts and removes this noise to generate the *edited image*, I_e. The *reverse instruction*, C_r, is then applied, using another sampled noise level $\hat{t}$, to recover the original image, I_r. This *single-step training approach* ensures that ERC is learned efficiently across varying noise scales, improving robustness across diverse edit scenarios.

By combining these three components—Text and Image Direction Alignment, Attention Map Alignment, and Reconstruction Alignment—our framework enforces ERC on real-image datasets *without requiring triplets of input-edited images with edit instructions*. This enables our approach enables to scale effectively beyond synthetic datasets, making it a practical solution for real-world instruction-based image editing.

4.1.2. Loss Functions

To enforce ERC, we introduce loss terms that guide editing and reconstruction. Each training iteration processes a *single-step training* sample consisting of an input image, an edit instruction, a reverse instruction, and captions.

CLIP Direction Loss. This loss ensures that the transformation applied to the image aligns with the text instruction in CLIP's semantic space [12]. Given the CLIP embeddings of the input image (E_{I_i}), edited image (E_{I_e}), input caption (E_{T_i}), and edited caption (E_{T_e}), the loss is defined as:

Table 1. **Reverse Instruction Generation.** Our method generates reverse instructions for the IP2P dataset, eliminating the need for ground-truth edited images. Additionally, edit instructions, edited captions, and reverse instructions are generated for CC3M and CC12M datasets—denoted as CCXM. The texts are generated by LLMs such as GEMINI, and GEMMA2.

	Input Caption	Edit Instruction	Edited Caption	Reverse Instruction
IP2P	A man wearing a denim jacket	make the jacket a rain coat	A man wearing a rain coat	make the coat a denim jacket
	A sofa in the living room	add pillows	A sofa in the living room with pillows	remove the pillows
	...	...	...	...
CCXM	Person on the cover of a magazine	make the person a cat	Cat on the cover of the magazine	make the cat a person
	A tourist rests against a concrete wall	give him a backpack	A tourist with a backpack rests against a concrete wall	remove his backpack
	...	...	...	...

$$\mathcal{L}_{\text{CLIP}} = 1 - \cos\left(E_{I_{\text{e}}} - E_{I_{\text{i}}}, E_{T_{\text{e}}} - E_{T_{\text{i}}}\right) \tag{1}$$

By aligning the direction of change in image space with the transformation described in text space, this loss ensures that modifications accurately reflect the intended semantic edits.

Attention Map Alignment Loss. To ensure that the same regions of the image are modified during both forward and reverse edits, we introduce an attention map alignment loss. Let $A_f^{(i)}$ and $A_r^{(i)}$ represent the cross-attention maps from the i-th layer of the U-Net model during the forward and reverse edits, respectively. The loss is defined as:

$$\mathcal{L}_{\text{attn}} = \sum_i \left\| A_f^{(i)} - A_r^{(i)} \right\|_2 \tag{2}$$

This loss ensures the model focuses on the same image regions during both forward and reverse edits—a key requirement for ERC. We compute it across U-Net's down and up layers at random timesteps. By encouraging consistent cross-attention maps, the loss promotes spatial coherence across noise levels, as fixed instructions tend to yield stable attention over edited regions (see Supp. Sec. 4.2).

CLIP Similarity Loss. This loss encourages the edited image to remain semantically aligned with the provided textual instruction. It is calculated as the cosine similarity between the CLIP embeddings of the edited image ($E_{I_{\text{e}}}$) and the edited caption ($E_{T_{\text{e}}}$):

$$\mathcal{L}_{\text{sim}} = 1 - \cos(E_{I_{\text{e}}}, E_{T_{\text{e}}}) \tag{3}$$

This loss ensures that the generated image aligns with the intended edits, maintaining semantic alignment with the provided instructions.

Reconstruction Loss. To guarantee that the original image is recovered after the reverse edit, we employ a reconstruction loss consisting of a pixel-wise loss and a CLIP-based semantic loss. The total reconstruction loss is defined as:

$$\mathcal{L}_{\text{recon}} = \|I_{\text{i}} - I_{\text{r}}\|_2 + [1 - \cos(E_{I_{\text{i}}}, E_{I_{\text{r}}})] \tag{4}$$

This loss ensures that the model can accurately reverse edits and return to the original image when the reverse instruction is applied, enforcing ERC by minimizing differences between the input and reconstructed images.

Total Loss. Overall, all losses are applied to a *single-step noise prediction*, making training more efficient. It is defined as a weighted combination of the individual losses:

$$\mathcal{L}_{\text{ERC}} = \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{sim}}\mathcal{L}_{\text{sim}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} \tag{5}$$

where λ_{CLIP}, λ_{attn}, λ_{sim}, and λ_{recon} are parameters that control the relative contributions of each loss term. The weights are selected through validation-based grid search to minimize $\mathcal{L}_{\text{ERC}}$ on a held-out set. Starting from the base configuration with $\mathcal{L}_{\text{CLIP}}$ and $\mathcal{L}_{\text{recon}}$, we found that $\mathcal{L}_{\text{sim}}$ enables freer edits by encouraging stronger image-text alignment, while $\mathcal{L}_{\text{attn}}$ improves localization by focusing on relevant regions during both forward and reverse edits.

4.2. Training Data

To enable ERC training on datasets with image and edit instructions [5], we leverage Large Language Models (LLMs), such as GEMMA2 [43] and GEMINI [42], to automatically generate reverse edit instructions. These LLMs provide an efficient and scalable solution for obtaining reverse instructions with minimal cost and effort [5]. We use GEMINI to enrich the IP2P dataset with reverse instructions based on the input caption, edit instruction, and corresponding edited caption. To improve model performance, we employ few-shot prompting during this process, enabling the generation of reverse instructions without the need for manually paired datasets, which significantly enhances scalability. The reverse instructions generated by the LLM aim to revert the edited image to its original form (see Tab. 1 - IP2P section). Using the enriched dataset with reverse instructions (see Tab. 1, IP2P section), we fine-tune GEMMA2 [43], to generate an edit instruction, edited caption, and reverse instruction given an input caption. We use

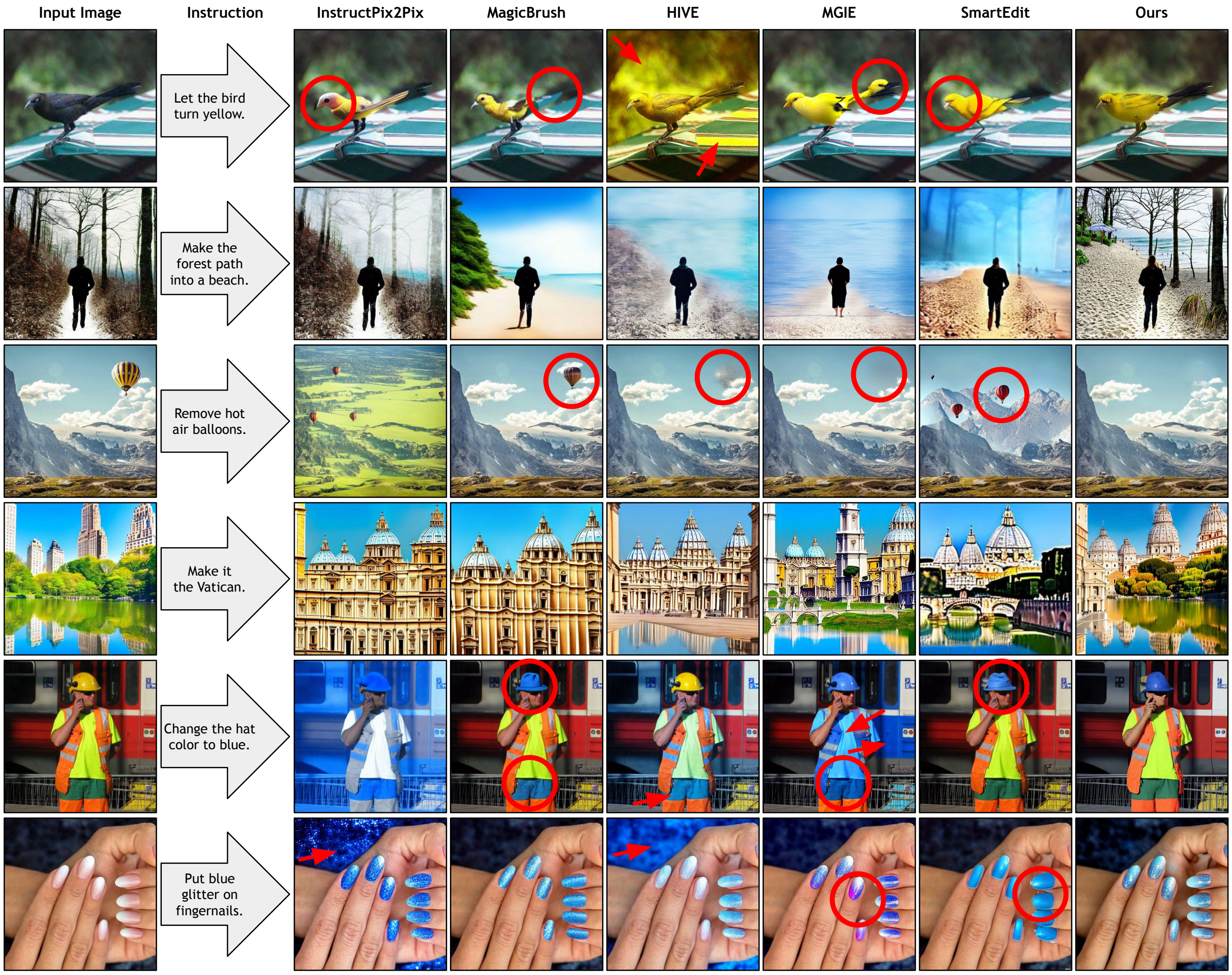


Figure 4. **Qualitative Examples.** UIP2P performance is shown across various tasks and datasets, compared to InstructPix2Pix, MagicBrush, HIVE, MGIE, and SmartEdit. Our method demonstrates either comparable or superior results in terms of accurately applying the requested edits while preserving visual consistency. Red circles and arrows indicate drastic problems during the image editing.

this fine-tuned model to allow training on image-caption paired datasets such as CC3M and CC12M [7, 37], generating forward and reverse edits along with corresponding edited captions (see Tab. 1, CCXM section).

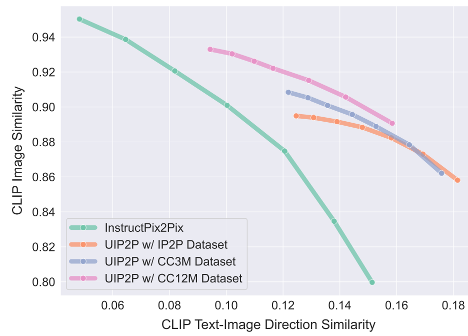
5. Experiments

Datasets. We generate datasets consisting of forward and reverse instructions, as detailed in Sec. 4.2. For the initial experiments, we use the InstructPix2Pix dataset [5], which provides generated image-caption pairs along with edit instructions. We further extend our experiments to real-image datasets, including CC3M [37] and CC12M [7], for which we generate eight possible edits per image-caption pair. This increases diversity in the editing tasks, exposing the model to a wide range of transformations and enhancing its ability to generalize across different types of edits and real-world scenarios. We then also conduct experiments on the HQ-Edit dataset [19], which already contains reverse instructions and higher-quality images, though they are still generated by off-the-shelf multimodal models [1, 3].

Baselines. We evaluate our method by comparing it against several models. The primary baseline is InstructPix2Pix [5], a supervised method that relies on ground-truth edited images during training. To demonstrate the advantages of our unsupervised approach, we train and test both IP2P and our model on the same datasets, with our method not using the ground-truth edited images while training. We also compare our method with other instruction-based editing models, including MagicBrush [51], HIVE [52], MGIE [10], SmartEdit [18], and HQ-Edit [19]. These comparisons allow us

Settings	Methods	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
Single-turn	HIVE [52]	0.1092	0.0341	0.8519	0.7500	0.2752
	InstructPix2Pix [5]	0.1122	0.0371	0.8524	0.7428	0.2764
	UIP2P w/ IP2P Dataset	0.0722	0.0193	0.9243	0.8876	0.2944
	UIP2P w/ HQ-Edit Dataset	0.0709	0.0190	0.9251	0.8882	0.2932
	UIP2P w/ CC3M Dataset	0.0680	0.0183	0.9262	0.8924	**0.2966**
	UIP2P w/ CC12M Dataset	**0.0619**	**0.0174**	**0.9318**	**0.9039**	0.2964
Multi-turn	HIVE [52]	0.1521	0.0557	0.8004	0.6463	0.2673
	InstructPix2Pix [5]	0.1584	0.0598	0.7924	0.6177	0.2726
	UIP2P w/ IP2P Dataset	0.1104	0.0358	0.8779	0.8041	0.2892
	UIP2P w/ HQ-Edit Dataset	0.1060	0.0344	0.8792	0.8081	0.2897
	UIP2P w/ CC3M Dataset	0.1040	0.0337	0.8816	0.8130	**0.2909**
	UIP2P w/ CC12M Dataset	**0.0976**	**0.0323**	**0.8857**	**0.8235**	0.2901

(a) **Zero-shot Quantitative Comparison on MagicBrush [51] test set.** Instruction-based editing methods that are not fine-tuned on MagicBrush are presented. In the multi-turn setting, target images are iteratively edited from the initial images.



(b) **Evaluation on the IP2P test dataset.** UIP2P outperforms IP2P in both CLIP image similarity and CLIP text-image similarity metrics, demonstrating better visual fidelity and instruction alignment.

Figure 5. Evaluation on MagicBrush and IP2P test datasets.

to evaluate how effectively our model handles diverse and complex edits without needing for editing methods to generate ground-truth edited images or human-annotated data.

Implementation Details. Our method, UIP2P, fine-tunes the SD-v1.5 checkpoint [34] without any pre-training on triplet datasets. While we retain the IP2P architecture (*Edit Model*, see Fig. 3), our approach employs different training objectives, focusing primarily on enforcing the ERC. Specifically, we leverage the CLIP ViT-L/14 model, already integrated into SD-v1.5, to compute the losses. Unlike supervised methods that begin training by adding noise to the ground-truth edited image, our method starts by adding noise to the input image. This design choice has significant implications, *e.g.*, it helps bias the model toward preserving the structure and details of the original image, ensuring that edits remain consistent with the input. Our method trains in a single denoising step, predicting only one step per iteration to ensure efficiency. UIP2P is trained using the AdamW optimizer [23] with a batch size of 256 over 11K iterations. The base learning rate is set to 1e-05. All experiments are conducted on 8 NVIDIA A100 GPUs, with loss weights set as $\lambda_{\text{CLIP}} = 1.0$, $\lambda_{\text{attn}} = 0.5$, $\lambda_{\text{sim}} = 1.0$, and $\lambda_{\text{recon}} = 1.0$. The best configuration is selected based on the validation loss of $\mathcal{L}_{\text{ERC}}$.

5.1. Qualitative Results

We compare UIP2P with state-of-the-art methods, including InstructPix2Pix [5], MagicBrush [51], HIVE [52], MGIE [10], and SmartEdit [18], on various datasets [5, 38, 39, 51]. The tasks include color modifications, object removal, and structural changes. UIP2P consistently produces high-quality edits, applying transformations accurately while maintaining visual coherence. For example, in "let the bird turn yellow," UIP2P provides a more natural color change while preserving the bird's shape. Similar improvements are observed in tasks like "remove hot air balloons" and "change hat color to blue." These results highlight UIP2P's ability to perform diverse edits, achieving successful and localized image modifications (see Fig. 4).

5.2. Quantitative Results

5.2.1. IP2P Test Dataset

We evaluate our method on the IP2P test split, containing 5K image-instruction pairs. Following [5], we use CLIP image similarity for visual fidelity and CLIP text-image similarity to assess alignment with the instructions. Higher scores in both metrics indicate better performance (upper right corner) by preserving image details (image similarity) and effectively applying the edits (direction similarity). As shown in Fig. 5b, UIP2P outperforms IP2P across both metrics. In these experiments, the text scale s_T is fixed, while the image scale s_I varies from 1.0 to 2.2.

5.2.2. MagicBrush Test Dataset

The MagicBrush test split contains 535 sessions (source images for iterative editing) and 1053 turns (individual editing steps). It uses L1 and L2 norms for pixel accuracy, CLIP-I and DINO embeddings for image quality via cosine similarity, and CLIP-T to ensure alignment with local text descriptions. As seen in Fig. 5a, UIP2P performs the best for both single- and multi-turn settings. It is important to be noted that HIVE utilizes human feedback on edited images to understand user preferences and fine-tunes IP2P based on learned rewards, aligning the model more closely with human expectations. Figure 5a shows that adding inverse instructions to IP2P improves performance, while training on HQ-Edit further enhances results due to its high-quality/resolution images and built-in reverse instructions. However, its smaller size (≈100K) and specialized collection pipeline limit its impact. In contrast, leveraging large-scale real-image datasets like CC3M and CC12M enables scalable training without labor-intensive data collection, achieving the best overall performance. See Appendix for more quantitative comparison.

5.2.3. User Study

We conduct a user study on the Prolific platform [30] with 50 participants to evaluate six methods—IP2P, MagicBrush, HIVE, MGIE, SmartEdit, and UIP2P—on 30 randomly sampled image-edit instructions from diverse datasets [5, 38, 39, 51]. Participants compare images from randomly paired methods, selecting the better-performing one across six pairs per question, yielding *18,000* total votes—9,000 for *edit success* and 9,000 for *localization accuracy*. In two runs, they first choose the model best following the instruction, *Successful*, then the one applying edits only to the intended area, *Localized*. To ensure robust evaluation, we use *Elo rating system* [9], a widely adopted ranking method in LLM and generative model benchmarking [4, 54]. By enabling continuous pairwise comparisons, Elo accounts for varying instruction difficulty and produces a relative ranking. Table 2 shows that UIP2P achieves the highest *Elo score* and adjusted *Win rate* across both criteria. See Appendix for more details.

Table 2. **User study.** Results comparing six methods based on Elo scores and top performer ratios for edit success and localization accuracy. UIP2P achieves the highest scores in both criteria.

	Successful		Localized	
Models	Elo score	Win rate	Elo score	Win rate
IP2P	1395	12%	1301	12%
MagicBrush	1482	16%	<u>1601</u>	<u>18%</u>
HIVE	1465	15%	1471	14%
MGIE	<u>1584</u>	<u>18%</u>	1529	17%
SmartEdit	1543	17%	1440	16%
UIP2P (Ours)	**1681**	**22%**	**1659**	**23%**

5.3. Ablation Study

Scalability. We train our method on IP2P, HQ-Edit, CC3M, and CC12M to assess its scalability with dataset size and quality. As shown in Fig. 5a, performance improves with larger and higher-quality datasets, with notable gains from IP2P to CC12M. This highlights our method's ability to leverage extensive data for better fidelity and localization. See Appendix for the discussion on runtime of our method.

Loss functions. We conduct a zero-shot evaluation on the MagicBrush test set (single-turn) to assess the contribution of each loss function, see Tab. 3 - Loss. Starting with the base configuration, containing $\mathcal{L}_{CLIP}$ and $\mathcal{L}_{recon}$, we observe moderate performance across the same metrics. Adding $\mathcal{L}_{sim}$ loss allows the model to perform edits more freely, as the *Base* without it tends to create outputs similar to the input image. Finally, $\mathcal{L}_{attn}$ enhances the model's focus on relevant regions and ensures that the region will remain consistent between the forward and reverse processes.

Attention Map Alignment Loss vs. Timestamp t. We compare the impact of using random timestamps t versus same timestamps t between forward and reverse instructions. As shown in Tab. 3 - Attn, using random t consistently outperforms the fixed approach across all metrics. This suggests that introducing variability in t during training leads to a more consistent cross-attention, resulting in better edit localization and overall performance. The rationale is that using different noise levels enhances robustness by exposing the model to varied denoising conditions during forward and reverse edits, while the attention consistency loss ensures the model maintains focus on the same regions regardless of the noise level. As also investigated in [14, 40], maintaining consistent attention maps across different timestamps enables more precise editing.

Table 3. **Ablation studies.** Adding additional loss functions to the base setup improves performance on the MagicBrush benchmark. Additionally, using random timestamps between forward and reverse instructions leads to better edit localization and overall performance.

	Ablation	L1↓	L2↓	CLIP-I↑	DINO↑	CLIP-T↑
	IP2P [5]	0.112	0.037	0.852	0.743	0.276
Loss	Base	0.117	0.032	0.878	0.806	**0.309**
	+ $\mathcal{L}_{sim}$	0.089	0.024	0.906	0.872	0.301
	+ $\mathcal{L}_{attn}$	**0.062**	**0.017**	**0.932**	**0.904**	0.296
Attn	Same t	0.074	0.022	0.921	0.889	**0.299**
	Random t	**0.062**	**0.017**	**0.932**	**0.904**	0.296

6. Conclusion

In this work, we introduce UIP2P, a novel unsupervised approach for instruction-based image editing that eliminates the need for paired supervision through the ERC. By enabling training on real-image datasets, our approach enhances scalability while maintaining high-fidelity edits. Through a single-step training paradigm, we achieve efficient and precise edits, outperforming existing methods. While our method advances instruction-based editing, future improvements in reverse instruction generation, spatial guidance, and diffusion model efficiency can further enhance performance. Additionally, incorporating perceptual or adversarial losses could improve photorealism, addressing cases where optimizing for CLIP alignment may compromise visual quality in challenging edits. As models continue to evolve, our approach stands to benefit from these advancements, further improving edit precision and inference speed. Overall, our work unlocks new possibilities for large-scale training in instruction-based image editing without explicit supervision. We encourage future research to build upon this foundation, refining both model design and training strategies.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 6

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18187–18197. IEEE, 2022. 1

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2, 6

[4] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*, 2023. 8

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7, 8

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 6

[8] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[9] Arpad Elo. *The Rating of Chess Players, Past and Present*. Batsford London, 1978. 8

[10] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 1, 2, 3, 6, 7

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 1

[12] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4):141:1–141:13, 2022. 3, 4

[13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023. 3

[14] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6986–6996, 2024. 8

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *CoRR*, abs/2208.01626, 2022. 1, 2, 3

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[18] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024. 1, 3, 6, 7

[19] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Cihang Xie, and Yuyin Zhou. HQ-edit: A high-quality dataset for instruction-based image editing. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 6

[20] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 3

[21] 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. *CoRR*, abs/2210.09276, 2022. 1

[22] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 2, 3

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[24] Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11451–11461. IEEE, 2022. 1

[25] 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 *International Conference on Learning Representations*, 2022. 1, 3

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *CoRR*, abs/2211.09794, 2022. 3

[27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, 2022*, pages 16784–16804. PMLR, 2022. 1

[28] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[29] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[30] prolific. Prolific. `https://www.prolific.com/`, 2025. Accessed: 2025-03-01. 8

[31] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, 2021*, pages 8748–8763. PMLR, 2021. 4

[32] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022*, pages 10674–10685. IEEE, 2022. 1, 3, 7

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 1

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022. 1

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[38] Jing Shi, Ning Xu, Trung Bui, Franck Dernoncourt, Zheng Wen, and Chenliang Xu. A benchmark and baseline for language-driven image editing. In *Computer Vision - ACCV 2020 - 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 - December 4, 2020, Revised Selected Papers, Part VI*, pages 636–651. Springer, 2020. 7, 8

[39] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 7, 8

[40] Enis Simsar, Alessio Tonioni, Yongqin Xian, Thomas Hofmann, and Federico Tombari. Lime: Localized image editing via attention regularization in diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 222–231, 2025. 8

[41] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 2, 3

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 5

[43] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 5

[44] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[45] Qian Wang, Biao Zhang, Michael Birsak, and Peter Wonka. Mdp: A generalized framework for text-guided image editing by manipulating the diffusion path, 2023. 3

[46] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1

[47] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023. 2, 3

[48] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 1900–1910, 2023. 3

[49] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. Cyclenet: Rethinking cycle consistent in text-guided diffusion for image manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[50] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 1

[51] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 6, 7, 8

[52] Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, Caiming Xiong, and Ran Xu. HIVE: harnessing human feedback for instructional visual editing. *CoRR*, abs/2303.09618, 2023. 1, 2, 3, 6, 7

[53] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37:3058–3093, 2024. 3

[54] Lianmin Zheng, Ying Sheng, Wei-Lin Chiang, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Chatbot arena: benchmarking llms in the wild with elo ratings. *LMSYS Org. URL: https://lmsys. org/blog/2023-05-03-arena [accessed 2024-04-16]*, 2023. 8

[55] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017. 2