Semantically Aligned Gradient-Driven Context-Preserving Image Editing

Chiranjeev Chiranjeev[1], Muskan Dosi[1], Mayank Vatsa[1], and Richa Singh[1]

Indian Institute of Technology Jodhpur, India
{chiranjeev.1,dosi.1,mvatsa,richa}@iitj.ac.in
https://github.com/IAB-IITJ/IABEdit

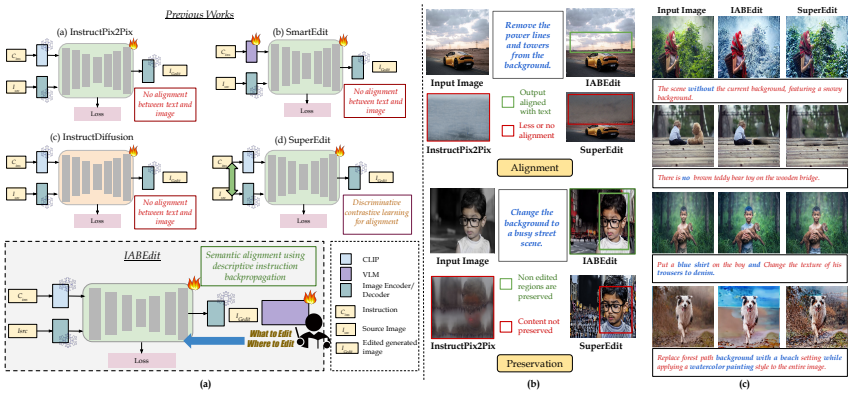

Fig. 1: (a) Existing editing models rely on static textual features, causing weak alignment and unwanted changes. (b) IABEdit improves both alignment and preservation through gradient-enabled semantic supervision. (c) Examples show IABEdit's stronger semantic grounding and localized edits compared to prior methods.

Abstract. Instruction-guided image editing has a training-time blind spot. Generative editors are never required to semantically verify whether their outputs actually satisfy the instruction. Supervision stops at reconstruction and input textual-level conditioning. This produces incomplete edits, spatial spillover, and poor localization. We present *IABEdit*, a model-agnostic framework that embeds differentiable semantic verification into training. A frozen vision-language model extracts spatially-aware descriptors from the ground-truth edit. A trainable aligner then reproduces them from the generated output. The residual between the two becomes a gradient that teaches the generator both what to edit and where, with no inference-time VLM cost. IABEdit is compatible with diverse backbones, including U-Net (Stable Diffusion) and MMDiT (FLUX), without altering their inference pipelines. On MagicBrush, it improves structural fidelity by +3.49 DINO-I over the best diffusion baseline and +1.26 over the best overall baseline, while remaining competitive on instruction alignment. It also achieves state-of-the-art instruction adherence performance on RealEdit and EMU Edit benchmarks based on

embedding-based metrics. Most consequentially, on the D-LORD surveillance benchmark, it surpasses the proprietary Gemini agent by +5.13 DINO-P under heavy occlusion, where preserving identity is hardest. This shows that gradient-aligned VLM distillation holds up under real-world-like surveillance and occlusion conditions. Human and GPT-4o evaluations confirm perceptually precise, well-localized edits.

Keywords: Image Editing · Semantic Alignment · VLMs

1 Introduction

Recent advances in generative image modeling, including diffusion models [8, 13, 23, 30] with U-Net and transformer-based generators, and flow-matching frameworks such as FLUX [18] have enabled high-quality and flexible image manipulation. Instruction-guided image editing extends these capabilities by allowing users to modify images through natural language, supporting tasks such as object removal, attribute modification, scene transformation, and semantic style transfer. Earlier approaches relied on manually specified spatial masks, limiting scalability and usability. Modern generative models [9, 18, 25, 27–29] instead condition directly using textual instructions, enabling language-based editing.

Despite their realism, existing instruction-guided editors are weak in instruction–edit alignment as shown in Figure 1(b), and share a fundamental limitation: the training objective does not explicitly verify whether the generated image semantically satisfies the intended edit. While instructions influence the generation process, there is typically no mechanism that measures alignment between the desired transformation and the produced output beyond reconstruction or token-level supervision. Crucially, the model is never required to verify whether the progressively denoised image actually satisfies the instruction. As a result, this design leads to systematic failure modes: (1) incomplete semantic realization (under-editing), (2) spatial mislocalization of edits, and (3) unintended modification of non-target regions (over-editing). The underlying issue is not architectural - it is the absence of explicit semantic supervision that enforces instructions.

We argue that instruction-guided editing should be formulated as a semantic alignment problem rather than purely a conditioning problem. To this end, we propose ***IABEdit***: *Instruction-Aligned Backpropagation Image Editing*. IABEdit introduces differentiable semantic verification into the training of generative image editors. Instead of relying solely on textual conditioning for obtaining intended edits, we incorporate high-capacity vision-language reasoning to define what a successful edit should represent at a semantic level. The overview of IABEdit is shown in Figure 1(a).

This semantic supervisor evaluates whether the generated image reflects the intended transformation, capturing both global changes and localized attributes. By making this evaluation differentiable, the generative model is guided by an explicit measure of instruction satisfaction during optimization. Editing thus becomes a process of minimizing the semantic discrepancy between the intended outcome and the generated image. This training paradigm encourages the model

to learn not only what must change, but also what must remain preserved, resulting in improved localization, stronger instruction adherence, and better structural consistency. IABEdit facilitates better alignment using supervision.

Importantly, IABEdit operates as an architecture-agnostic training-time supervision layer. It can be integrated into diffusion models, transformer-based generators, including MMDiTs in flow-matching frameworks, without altering their inference pipelines. Rich vision-language reasoning supervision is used only during training; at deployment (testing phase), editing remains a standard generative process with no additional computational overhead. By reframing instruction-guided editing as semantic alignment optimization, IABEdit establishes a principled and scalable framework for semantically grounded image editing. Across diverse benchmarks and generative backbones, we demonstrate that enforcing semantic consistency during training yields more faithful, localized edits, independent of the underlying architecture. Our key contributions are:

1. **Differentiable Semantic Reasoning-based Verification:** We introduce a gradient-based training objective that backpropagates VLM-derived semantic descriptors through the generative backbone, explicitly teaching both *what* to edit and *where* to edit. To the best of our knowledge, IABEdit is among the first methods to embed textual semantic reasoning-oriented instruction verification directly into the diffusion and flux training loop.
2. **Model-Agnostic, *VLM-free* Inferencing:** IABEdit operates as a training-time supervision layer requiring no changes to inference pipelines, making it directly applicable to UNet, transformers in flow-matching architectures with no additional runtime overhead.
3. **State-of-the-Art Results Across Domains:** IABEdit demonstrates improvements across diverse editing tasks: Scene-level editing and Identity-sensitive Facial editing. IABEdit achieves leading performance on RealEdit, MagicBrush, EMU Edit and D-LORD benchmarks.

1.1 Related Work

Earlier image manipulation approaches relied on mask-based editing [5, 32, 33], which provided explicit localized control but required manual spatial annotations. Instruction-based editing methods instead use image generation models [2,10,11,19] to embed textual conditions, typically following a triplet setup of source image, instruction, and target edited image. Many earlier works [12, 19] generated training data by rectifying instructions with LLMs and producing edited images via diffusion models. However, current T2I models apply edits in unintended regions, resulting in noisy supervision. To improve quality, [10,19,34] introduced high-quality editing datasets through image filtering [17]. Despite these efforts, most methods still rely on static textual CLIP embeddings [26], used purely as external conditioning. This often leads to misalignment between fine-grained instructions and edited regions. Accurate localization of instruction-relevant elements (e.g., power lines) is crucial for targeted edits (Figure 1b).

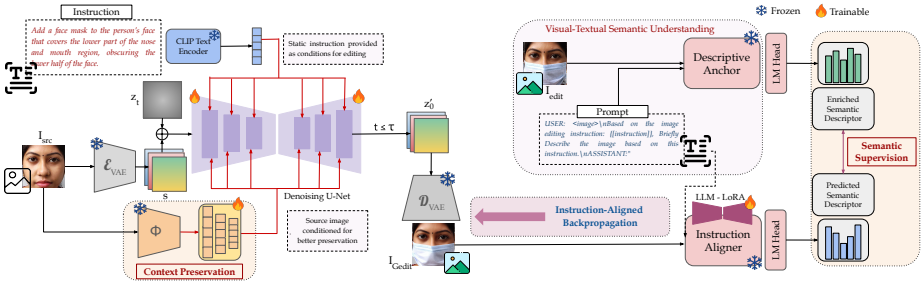

Fig. 2: IABEdit enables instruction-based image editing through a semantic-supervised distillation mechanism. A frozen Descriptive Anchor extracts rich guidance from the instruction and ground-truth edit, while an Instruction Aligner learns to align the generated image with this guidance. The alignment is enforced via backpropagation, guiding the diffusion model toward faithful and controllable edits.

To improve semantic understanding, recent methods [12, 35] incorporate reward information into text conditions. [15] replace CLIP with VLMs [21, 24] to obtain richer textual understanding. However, these approaches still treat the instruction only as a prior conditioning signal in the diffusion process, without providing active or corrective understandings, limiting fine-grained alignment. They also incur high inference cost due to VLM-based conditioning. SuperEdit [19] addresses misalignment through rectified prompts and contrastive supervision, building a discriminative link. Yet, this operates mainly at the prompt level, leaving instruction semantics out of the generation process. Overall, existing methods remain constrained by static conditioning, weak localization, and costly inference, underscoring the need for methods that achieve deeper, more reliable instruction-to-image alignment.

2 IABEdit: Instruction-Aligned Editing

We propose IABEdit, which provides a balance between context preservation and semantic editing. As illustrated in Figure 2, our approach consists of two components: (1) image context conditioning, which maintains structural fidelity through encoder and conditioning-level preservation, and (2) editing through gradient-enabled alignment feedback, which provides semantic supervision during training to enforce instruction-consistent modifications.

2.1 Editing using context as condition

We formalize successful image editing as a balance between two competing objectives: faithful realization of the instruction and preservation of the original visual context. In practice, many editing models rely primarily on static textual conditioning while using the input image only as a spatial prior, which conditions

what needs to be edited but does not focus on what needs to be preserved. Thus, edits may propagate beyond the intended regions or alter unrelated content.

To explicitly enforce context preservation, IABEdit introduces structured conditioning that separates transformation cues from context cues. The generation process is guided by three complementary signals: (i) a textual instruction embedding c_{ins} that specifies the desired modification, (ii) a source-image embedding c_{src} that encodes high-level visual attributes and context information, and (iii) the latent representation s of the source image to retain spatial structure. By decoupling edit intent from contextual cues, this formulation encourages precise modifications while minimizing unintended changes.

The instruction embedding c_{ins} is obtained from a text encoder, while the source-image embedding c_{src} is derived from visual features of the input image and projected into a conditioning space using a multi-layer linear projector (MLP). These signals are injected into the generative model (e.g., via cross-attention) to guide the editing process. For clarity, we present the diffusion-based formulation in the main paper, while flow-based variants are provided in the supplementary material:

$$\begin{aligned} c_{ins} &= \mathrm{CLIP}(\mathit{instruction}), \quad c_{src} = MLP(\phi(I_{src})), \\ z_t &= \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \\ \mathcal{L}_N &= \mathbb{E}\left[\left\|\epsilon - \epsilon_\theta\left(z_t \oplus s,\ c,\ t\right)\right\|_2^2\right]. \end{aligned} \tag{1}$$

Here, z_t denotes the noisy latent at timestep t, c denotes all conditioning inputs, $c = \{c_{\mathrm{ins}}, c_{\mathrm{src}}\}$. and $\mathcal{L}_N$ is the standard denoising objective. During training, we apply dropout to the instruction embedding c_{ins} and spatial latent s to enable classifier-free guidance. In contrast, the source-image embedding c_{src} is always preserved, as it encodes identity and contextual cues that must remain consistent—particularly in tasks such as face editing.

Importantly, context preservation and edit fulfillment play fundamentally different roles during generation. The contextual cues (s) of the source image remain constant throughout the denoising trajectory, and therefore can be represented as structurally conditioning latents. In contrast, the degree to which the instruction has been satisfied evolves dynamically at each iteration of generation. The regions requiring modification depend on the current state of the image. Static conditioning alone cannot account for this evolving semantic discrepancy.

This observation motivates the need for a dynamic, output-dependent supervision signal that continuously evaluates how much editing is still required. IABEdit introduces such a mechanism through semantic feedback, enabling the model to adaptively refine edits while preserving context.

2.2 Semantic Alignment via Gradient-Enabled VLM Supervision

Conventional instruction-based editing methods encode instructions once via CLIP and inject them as static conditioning signals. This pre-conditioning approach suffers from two critical limitations: short CLIP embeddings lack expressiveness for complex instructions, and the model never verifies whether generated

outputs satisfy the instruction. These limitations manifest as systematic failures, as shown visually in Figure 3, where *adding a hat "above" rather than "on" the cat's head* indicates an abstract semantic understanding of the instruction. Further, the image editing *removes both the teddy bear and the child when instructed to remove only the bear*, which shows the lack of spatial localized understanding of the context. Thus, the static conditioning focuses on instruction tokens but lacks a mechanism to infer the underlying intent of the instruction, such as determining edit-remaining regions/tokens.

We address this limitation by leveraging the semantic reasoning capabilities of vision-language models (VLMs). A frozen VLM, Descriptive Anchor, serves as a semantic reference, extracting detailed descriptors from the ground-truth edited image to capture the intended transformation. A trainable VLM, Instruction Aligner, is then optimized to predict these descriptors from the generated output, effectively measuring the discrepancy between the realized and desired edits. The resulting semantic alignment loss produces gradients proportional to the residual discrepancy between the generated image and the intended edit. These gradients are backpropagated to the image generative model, providing output-dependent supervision that adaptively guides both *what* should be modified and *where* modifications should be localized.

Through this continuous semantic correction process, IABEdit transforms instruction following from static conditioning into iterative alignment, enabling the model to leverage VLM-level understanding during training while learning to perform precise and efficient edits. Importantly, this semantic feedback operates only during training; inference remains a standard generative process without additional computational overhead.

Ground-Truth Semantic Supervision: We employ a frozen vision-language model (VLM) to extract rich semantic representations from the ground-truth edited image I_{edit} conditioned on the instruction prompt P. Unlike simple captions, these descriptors encode both global transformations (e.g., "watercolor style applied throughout") and fine-grained localized attributes (e.g., "blue hat on the cat's head, covering the ears"). This captures spatial relationships, colors, and occlusions far beyond static CLIP embeddings. As they are automatically generated through image understanding, no additional annotations are required.

Learning Semantic Alignment: A trainable VLM, implemented via lightweight LoRA adaptation [14], is optimized to reproduce these semantic representations from the generated image $I_{G_{edit}}$. Matching the generated and reference semantics defines a differentiable alignment objective, enabling gradients to flow back into the image generative model and enforce instruction-consistent edits.

To obtain z'_0 without full diffusion sampling, we employ single-step denoising at small timesteps $t \leq \tau$. At these early timesteps, noisy latent z'_t closely approximates clean latent z_0 since noise level $\sqrt{1-\bar{\alpha}_t}$ is minimal:

$$z_0 \approx z'_0 = \frac{z'_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta((z'_t \oplus s), c_{ins}, c_{src}, t)}{\sqrt{\bar{\alpha}_t}} \tag{2}$$

This approximation [20] provides precise estimates while avoiding the memory overhead of 50+ DDIM steps per training batch.

Alignment-Relevant gradient flow: We train our dynamic semantic aligner via KL divergence minimization:

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} &= \frac{1}{B}\sum_{b=1}^{B}\mathrm{KL}\left(\mathrm{Softmax}\left(\frac{\mathbf{l}_f^{(b)}}{T}\right)\middle\|\mathrm{Softmax}\left(\frac{\mathbf{l}_t^{(b)}}{T}\right)\right) * T^2 \\ &= \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N} p_f^{(b)}(i)\left[\log p_f^{(b)}(i) - \log p_t^{(b)}(i)\right] * T^2\end{aligned} \tag{3}$$

where $\mathbf{l}_f^{(b)}$ and $\mathbf{l}_t^{(b)}$ denote the token-level logits from the frozen Descriptive Anchor and the trainable semantic aligner for the b-th sample over N tokens. A temperature T is used to soften the probability distributions, enabling the aligner to learn from fine-grained semantic discrepancies; the T^2 factor compensates for gradient scaling during distillation. The corresponding token probabilities are denoted as $p_f^{(b)}(i)$ and $p_t^{(b)}(i)$.

Crucially, spatial localization emerges through gradient flow. The alignment loss $\mathcal{L}_{\mathrm{D}}$ is computed from the generated image $I_{G_{edit}}$, which depends on the predicted latent z_0' and ultimately on the model's noise estimator ϵ_θ. Backpropagating $\nabla\mathcal{L}_{\mathrm{D}}$ therefore injects semantic discrepancy signals directly into the generative backbone, guiding updates toward regions that require modification to satisfy the instruction. Unlike static conditioning, which provides fixed semantic priors, IABEdit embeds semantic reasoning verification into the optimization objective itself. Instruction adherence is no longer assumed—it is explicitly minimized as part of the training landscape.

Joint Optimization of Context Preservation: Distillation gradients also update the MLP projector mapping $\phi(I_{src})$ to source embedding c_{src}. Rather than encoding generic features, the projector learns instruction-aware conditioning: emphasizing features in regions requiring edits while preserving unchanged areas. This joint optimization of textual (via distillation) and visual (via MLP) conditioning enables spatially precise edits.

This framework enables faithful instruction-guided editing across general scenes; however, certain domains such as face editing require additional constraints to preserve subject-specific identity characteristics during modification.

Identity Preservation for Face Editing: For face editing applications in surveillance and security contexts, maintaining biometric identity under occlusions (masks, sunglasses) is critical. While semantic distillation ensures instruction alignment, it does not explicitly constrain identity preservation. We therefore introduce an identity coherence loss $\mathcal{L}_C$ that minimizes cosine distance between facial embeddings:

$$\mathcal{L}_{\mathrm{C}} = \frac{1}{B}\sum_{b=1}^{B}\left(1 - \cos(\phi^{(b)}(I_{src}), \phi^{(b)}(\mathcal{D}_{\mathrm{VAE}}(z_0')))\right) \tag{4}$$

where ϕ is a pre-trained face recognition model (ArcFace [7]). This encourages retention of biometric features in the generated edit $\mathcal{D}_{\text{VAE}}(z_0')$ despite occlusions or appearance changes. By operating in the embedding space of a discriminatively trained face recognition model, this loss ensures that identity-critical features (facial structure, proportions) remain invariant even when instruction-guided edits add realistic occlusions that would otherwise distort facial geometry.

The complete loss function combines all components conditionally:

$$\mathcal{L}_{\text{IABEdit}} = \begin{cases} \lambda_N \cdot \mathcal{L}_N + \lambda_D \cdot \mathcal{L}_D + \lambda_C \cdot \mathcal{L}_C, & \text{if } t \leq \tau \\ \lambda_N \cdot \mathcal{L}_N, & \text{otherwise} \end{cases} \tag{5}$$

where λ_{N}, λ_{D}, and λ_{C} are scalar weights that balance the contributions of each loss term. τ refers to the timestep threshold, which is a hyper-parameter used to determine whether the generated edited image should be utilized for VLM fine-tuning and maintaining structural coherence of faces. Distillation ($\mathcal{L}_D$) and identity ($\mathcal{L}_C$) losses apply only at small timesteps ($t \leq \tau$) where generated outputs are available via single-step denoising. For general scene editing tasks, $\mathcal{L}_C$ is omitted, and the framework optimizes only $\mathcal{L}_N$ and $\mathcal{L}_D$.

3 Experimental Details

We present a detailed experimental setup for rigorously evaluating our editing mechanism with various benchmarks, a model-agnostic generative method and balanced evaluation criteria.

Datasets: For scene image editing tasks, we use Super-40K [19] as a training dataset. Evaluation is conducted on RealEdit [12], the MagicBrush [34], and the EMU Edit [31] test set to assess image preservation and instruction alignment. To assess the performance of the model in identity-sensitive applications such as surveillance, we further train and evaluate on the disjoint train-test sets of D-LORD [22] dataset [1], which includes facial images under occlusion.

Architecture and Implementation Details: The IABEdit framework is model-agnostic, can work with diffusion and flow-based generative models. Semantic supervision is provided using a LLaVA-based [21] VLM model. During inference, IABEdit functions as a lightweight diffusion-only model. It uses CLIP-based 77-token instruction embeddings (where instructions comprise 20-30 effective textual tokens) and conditional image features (CLIP for scene edits and ArcFace for facial edits) with a learned small set of MLP layers. To incorporate deeper semantic guidance with detailed descriptive context, we use LLaVA-based 1362-token (with 100 effective generated textual tokens) descriptors, enabling fine-grained edits aligned with instruction semantics. Additional implementation details, and training and testing protocols are provided in the supplementary.

[1] The dataset was obtained through: `https://dyslai.org/`

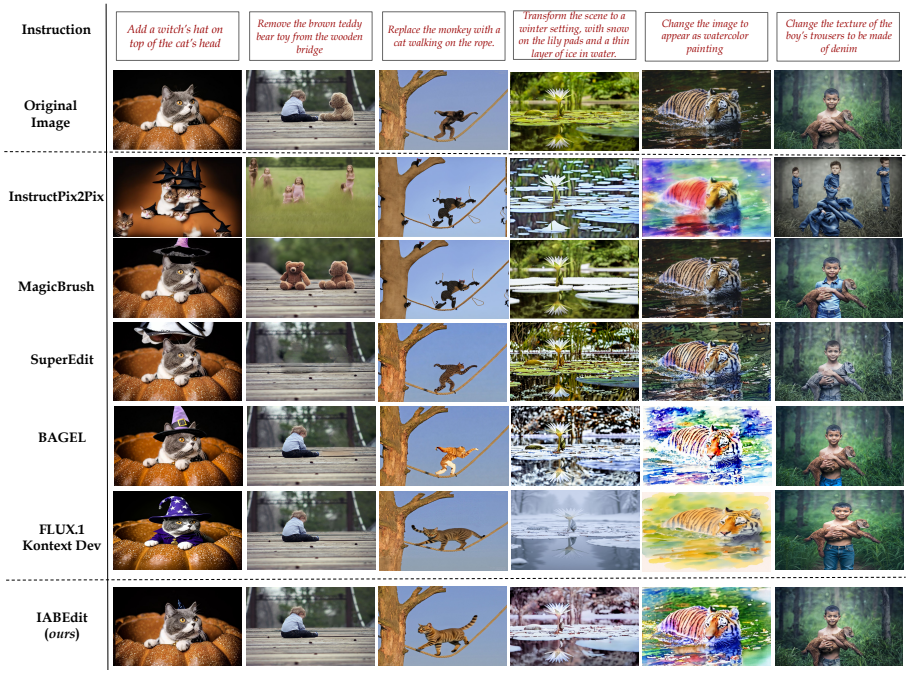

Fig. 3: Qualitative results on RealEdit dataset showing comparisons across various editing methods. The instruction adherence can be visualized at the fine-grained level. The non-targeted regions remain preserved, showcasing precise instruction-aligned editing generations. More visual results are available in supplementary.

Evaluation Metrics (Alignment and Preservation): IABEdit is evaluated in terms of quantitative metrics, agent-based assessments, as well as human evaluations. To assess localized edits and structural preservation, we use feature-level metrics: CS-P (Content Similarity Preserving) for measuring content retention of non-edited regions by computing CLIP-similarity between the source and edited image, CLIP-I for measuring similarity to the target-edited image, and CLIP-T for instruction alignment that leverages a pre-trained CLIP model. DINO-I [3] and L1 scores measure structural and pixel-level similarity to the ground truth edited image, while DINO-P is used to measure the facial identity structural preservation of the edited face with respect to source face. To capture the trade-off between instruction adherence and content preservation, we compute Harmonic Mean (HM) between CLIP-T and CS-P, offering a unified measure of balanced editing performance. This helps quantify whether an edit faithfully follows the instruction without significantly altering the non-editing regions. As traditional metrics often miss perceptual and semantic subtleties that humans easily identify [12,15,19], we incorporate MLLM agents like GPT-4o [17] to evaluate alignment, spatial precision, and realism, complemented by a user study grounded in the same criteria.

4 Results and Analysis

We evaluate IABEdit across multiple datasets and compare it with state-of-the-art instruction-based editing methods. While recent work has improved performance through dataset curation [16, 34] and instruction refinement [12, 19], a fundamental challenge remains: directly aligning textual semantics with precise spatial editing regions during training. As shown in Figure 3, IABEdit addresses this challenge through gradient-based semantic distillation, yielding superior instruction adherence and localization.

Instruction Alignment and Localized Edit Control: As shown in Figure 3, existing methods struggle with precise localization. For instance, SuperEdit removes not only the teddy bear but also the child (column 2), and adds denim texture to both the boy's trousers and the goat (column 5). Whereas an autoregressive multi-modal generative method, such as BAGEL, also fails to completely remove the teddy bear from the bridge *(zoom in for a better view)*, and also just changes the color of the trousers without changing their texture to denim. Similarly, a lack of understanding of the instructions' semantics is observed in recent FLUX.1 Kontext Dev method, it over-edits as it not only adds the hat on the cat's head, but it also puts a jacket on the cat (see Figure 3 column 1), similarly, it changes the trousers to denim but also increases the length of the boy's legs. These errors reflect weak coupling between instruction semantics and spatial edits. In contrast, IABEdit's gradient-based distillation enables fine-grained control, modifying only instruction-relevant regions. Quantitatively also, as shown in Tables 1a and 1b, IABEdit with diffusion achieves the highest CLIP-T scores among diffusion-based methods on RealEdit (27.60) and MagicBrush (30.97) benchmarks, which is superior to recent diffusion models (UltraEdit and SuperEdit) and perform competitive to computationally heavy flux-based (FLUX.1 Kontext Dev) and autoregressive (BAGEL) image editing methods. IABEdit also achieved a CLIP-I score of 92.41 on the MagicBrush test set, indicating strong alignment between the edited image and the target edited image, showing the controlled localized edit.

As shown in Table 1b, IABEdit achieves the best DINO-I score on MagicBrush (88.26), outperforming the strongest diffusion-based baseline (UltraEdit) by +3.49 points, and attains the lowest L1 distance (0.058), indicating sharper structural fidelity to the ground-truth edits. Moreover, GPT-4o score evaluations (Table 2 "Following" dimension) also rank IABEdit highly for instruction adherence (3.63), affirming its semantic precision in fine-grained edits.

Content and Structural Preservation: IABEdit excels at maintaining the integrity of non-instructed image regions, preserving both visual content and spatial structure. It can be observed from Figure 3, that recent state-of-the-art FLUX-based editing method (FLUX.1 Kontext Dev) does not preserve the context of the image as shown through the distortion in the background context behind the lily flowers (column 3, row 6), and in the 'watercolor-style edit', it fails to preserve the water region present in the original tiger image. Quantitatively, this is reflected by achieving high DINO-P and CS-P scores across datasets, indicating that the original image semantics are retained even after

Table 1: (a) Comparing performance of IABEdit (Diffusion+LLaVA) with other image editing methods on the RealEdit dataset. (b) Comparing performance of IABEdit (Diffusion+LLaVA) with other image editing methods on the MagicBrush benchmark.

(a)

Method	CS-P ↑	CLIP-T ↑	HM ↑
MagicBrush [34]	90.25	26.98	41.28
InstructDiffusion [11]	74.20	26.69	38.71
InstructPix2Pix [2]	75.76	27.00	39.78
Reward-InstructPix2Pix [12]	89.69	27.31	41.67
MGIE [10]	84.64	26.20	39.58
SmartEdit [15]	87.70	26.50	40.64
HQ-Edit [16]	66.32	27.09	38.21
UltraEdit [36]	84.91	27.50	41.39
SuperEdit [19]	**90.93**	26.01	40.24
BAGEL [6]	90.60	26.69	41.00
FLUX.1 Kontext Dev [18]	87.64	27.21	41.29
IABEdit (Ours)	89.65	**27.60**	**42.00**

(b)

Method	CLIP-I ↑	CLIP-T ↑	DINO-I ↑	L1 ↓
MagicBrush [34]	90.70	30.60	80.60	0.062
InstructDiffusion [11]	89.20	30.20	77.70	-
InstructPix2Pix [2]	85.40	29.20	69.80	0.112
Reward-InstructPix2Pix [12]	88.90	29.80	-	-
MGIE [10]	90.90	30.50	-	-
SmartEdit [15]	90.40	30.30	79.70	0.081
HQ-Edit [16]	69.08	25.88	51.62	0.243
UltraEdit [36]	90.47	30.86	84.77	0.066
SuperEdit [19]	90.50	30.30	80.20	0.106
BAGEL [6]	92.13	31.02	87.00	-
FLUX.1 Kontext Dev [18]	91.90	**31.28**	86.03	-
IABEdit (Ours)	**92.41**	30.97	**88.26**	**0.058**

Table 2: Performance comparison on the RealEdit dataset (having a sample size of 560 instruction-image pairs) across various instruction-based image editing methods. The evaluations are carried out using GPT-4o.

Method	Following ↑		Preserving ↑		Quality ↑		Overall ↑	
	Acc (%)	Score	Acc (%)	Score	Acc (%)	Score	Acc (%)	Score
MagicBrush [34]	51.00	2.90	70.00	3.85	50.00	3.67	57.00	3.47
InstructDiffusion [11]	52.00	2.87	54.00	3.17	45.00	3.58	50.30	3.21
InstructPix2Pix [2]	52.00	2.94	53.00	3.31	50.00	3.69	51.70	3.31
Reward-InstructPix2Pix [12]	63.00	3.39	58.00	3.43	54.00	3.80	58.30	3.54
MGIE [10]	40.00	2.43	45.00	2.79	38.00	3.35	41.00	2.86
SmartEdit [15]	64.00	3.50	66.00	3.70	45.00	3.56	58.30	3.59
SuperEdit [19]	**67.00**	3.59	**77.00**	4.14	**65.00**	**4.01**	**69.70**	**3.91**
IABEdit (Ours)	66.00	**3.63**	72.00	**4.19**	54.00	3.54	64.00	3.78

editing. Moreover, IABEdit achieves superior DINO-P scores, particularly on the D-LORD dataset, demonstrating its strength in preserving identity-specific structural features, especially in face editing tasks as compared to a proprietary image-generation Gemini-AI agent. The framework's use of cross-attention and latent conditioning enables localized edits without unintended modifications. Together, these results highlight IABEdit's ability to perform precise edits while preserving the non-edited regions of original image.

Balanced Editing: Preserving While Transforming: IABEdit optimally balances instruction adherence and content preservation, achieving the highest Harmonic Mean (HM) between CLIP-T and CS-P metrics (Table 1a). On RealEdit, IABEdit scores HM=42.00, significantly outperforming Reward-InstructPix2Pix (41.67), UltraEdit (41.39), SuperEdit (40.24), BAGEL (41.00), FLUX.1 Kontext Dev (41.29) and other baselines. We also performed the Wilcoxon signed-rank test with n=560 and observed significant results for all models with $p < 0.05$, whose detailed values are shown in the supplementary.

IABEdit's CS-P score of 89.65 is slightly lower than SuperEdit's 90.93, but this difference reflects successful editing rather than a limitation. When instructions require substantial changes, the edited image naturally diverges from the source. Many competing methods struggle with this balance: they either preserve too much content and under-edit (high CS-P but low CLIP-T), or they over-edit and introduce unintended changes (low CS-P with inconsistent CLIP-T). IABE-

Fig. 4: (a) Qualitative comparison of IABEdit (+UNet Diffusion) on MagicBrush test-set with other state-of-the-art editing methods. (b) Qualitative visualization of RealEdit test-set samples showcasing improvement of FLUX.1 method by training it with the IABEdit framework.

dit's gradient-based distillation provides precise spatial control, modifying only the regions specified by the instruction while leaving other areas unchanged.

Agent-Based and Human-Centric Evaluation: Table 2 reports GPT-4o evaluations on RealEdit across instruction following, content preservation, and image quality measures. *Score* measures fine-grained assessment on a 0-5 scale while *Accuracy* acts as a binary acceptability criterion. Although SuperEdit is more consistent at passing the binary baseline, IABEdit attains higher continuous scores in Instruction Following (3.63) and Content Preservation (4.19), indicating more precise edits while retaining non-target regions. IABEdit's perceptual quality score (3.54) is slightly lower than SuperEdit, but the overall score remains competitive at 3.78 (second-best). The quality-based comparisons are discussed in the supplementary material. Overall, it reveals a trade-off: SuperEdit offers robustness, while IABEdit provides higher-fidelity edits and preservation. On D-LORD facial edits, IABEdit attains an instruction-following score of 4.67 with 94.8% accuracy, comparable to Gemini-AI (4.53, 92.19%).

Comparison with Distinct Architectural Editing Methodologies: We compare IABEdit against fundamentally different architectures: FLUX.1 Kontext Dev [18] (flow matching) and BAGEL [6] (unified autoregressive multimodal mixture of transformers). Although these methods generate photorealistic output, they exhibit three failure modes (Figure 4a): *Keyword fixation:* focusing on isolated words ("fries", "car") rather than full instruction semantics. *Spatial misalignment:* inserting objects without respecting scene composition (e.g., overlaying a car on an existing bike). *Global distortion:* altering color/lighting beyond edited regions, and overshooting scale by adding "too large basil leaves" without respecting the image content's scale.

We observe that, even with a lightweight UNet-based diffusion method in the IABEdit framework, it performs better-balanced image editing (refer to Figure 4(a)–qualitative visualization and quantitative performance observed through Table 1a and Table 1b) as compared to the computationally heavy flux and autoregressive methods.

Model Agnostic IABEdit Framework: Table 1a and 1b show results with IABEdit using a UNet-based denoising backbone, which outperforms existing

diffusion-based methods such as SuperEdit, InstructPix2Pix and Reward InstructPix2Pix. Moreover, when IABEdit is trained over FLUX.1 model, it improves the image quality of the generated edited image, this property is inherited from FLUX's quality. As we noted that FLUX.1 is instruction's keyword fixated and does not understand the instruction's semantics, further it also does not focus on the image's context. We solve this problem by training FLUX.1 with the IABEdit framework, which improves its image-text semantic alignment and also preserves the image context in the generated edited image. After training FLUX.1 with IABEdit, it improves FLUX.1 on the RealEdit dataset by attaining a CS-P score of ***90.20***, CLIP-T score of ***27.78*** and performs balanced editing by achieving a HM score of ***42.28***. While FLUX.1 achieves a CS-P score of ***87.64***, a CLIP-T score of ***27.21*** and performs balanced editing by achieving a HM score of ***41.29***. All these scores achieved on IABEdit+FLUX.1 are significantly higher, as they attained a p-value$<$0.05. It can be observed that IABEdit+FLUX.1 improves the image context preservation score compared to the existing FLUX.1 model. Figure 4(b) presents qualitative results of FLUX.1+IABEdit. It can be observed that the context of the original image is preserved in IABEdit (+FLUX.1) while following the editing instruction, whereas FLUX.1 (Kontext Dev) destroys the image context. Therefore, IABEdit is a model-agnostic framework suitable for all kinds of denoising networks, including UNet in Stable Diffusion and MMDiTs in FLUX.1.

Results on the challenging EMU Edit benchmark [31] across diffusion and FLUX-based architectures are provided in the supplementary. IABEdit outperforms the strongest baseline by +1.41 in CLIP-based instruction following, +0.84 in target-caption alignment, and +9.98 in DINO-based structural preservation, demonstrating semantically accurate edits while preserving image structure.

Real-World Surveillance Scenarios: In real-world surveillance scenarios, instruction-based facial editing is particularly challenging due to realistic transformations like disguises (masks and sunglasses), while preserving identity. As shown in Table 3, IABEdit outperforms the Gemini-AI (2.5Pro) [4] and SuperEdit methods, achieving higher CS-P (79.93) and CLIP-T (33.10) scores. We also performed a Wilcoxon signed-rank test on the harmonic mean scores and achieved a p-value of < 0.05. An improvement of +5.13 DINO-P points over Gemini-AI (agent) method and +3.69 over SuperEdit (trained on D-LORD train set and evaluated on disjoint identity test set) indicates stronger preservation of structural identity. We have also evaluated the effectiveness in terms of face recognition performance. While the detailed results are included in supplementary, combining the generated edited facial samples with the real face dataset improved the recognition performance by 2.82%.

Table 3: Quantitative comparison with agent-based generation models on the real-world D-LORD dataset for facial editing.

Generation Technique	CS-P ↑	CLIP-T ↑	HM ↑	DINO-P ↑
SuperEdit [19]	<u>79.19</u>	22.96	35.38	<u>51.64</u>
Gemini-AI (Agent) [4]	78.67	<u>31.40</u>	<u>44.79</u>	50.20
IABEdit (Ours)	**79.93**	**33.10**	**46.74**	**55.33**

Ablation Study: ***(1) IABEdit's Component Ablation:*** Figure 5 presents the ablation results highlighting the direct impact of the proposed IABEdit method based on the loss components introduced in our method. The effective-

Fig. 5: Ablation results on the Real-Edit dataset.

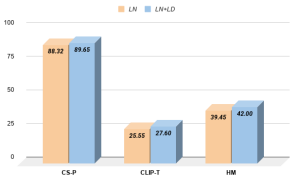

Table 4: Comparisons on RealEdit dataset using 3 protocols: (1) with and w/o LoRA-finetuning, (2) VLM models, and (3) with and w/o context preservation conditioning.

	CS-P (↑)	CLIP-T (↑)	HM (↑)
Without LoRA-finetuning	83.44	24.44	37.58
With LoRA-finetuning (IABEdit)	**89.65**	**27.60**	**42.00**
IABEdit (LLaVA-7B)	89.65	27.60	42.00
IABEdit (Qwen-VL 2.5-7B)	89.33	**27.91**	**42.27**
IABEdit (w/o context preservation)	88.91	25.90	40.19
IABEdit (w/ context preservation)	**89.65**	**27.60**	**42.00**

ness of our method for balanced editing is demonstrated by higher HM scores achieved when training the model with combined Denoising ($\mathcal{L}_N$) and Distillation ($\mathcal{L}_D$) loss functions. Incorporating both objectives results in +2.47 performance improvement compared to training with only Denoising loss ($\mathcal{L}_N$). This underscores the importance of the distillation component with gradient-enabled meaningful supervisory signals. It enhances the image generation model's ability to perform semantically aligned edits with natural language instructions.

(2) Impact of LoRA fine-tuning: Table 4 (first two rows) shows ablation results for highlighting the utilization of LoRA adapter for fine-tuning the predicted semantic VLM descriptor. Without LoRA adaptation, the Instruction Aligner operates on noisy intermediate denoised outputs and produces unstable semantic descriptors, resulting in inaccurate gradient updates from $\nabla\mathcal{L}_D$. LoRA adaptation significantly improves robustness to intermediate generations, enabling semantic supervision, while the frozen supervisor is a stable reference. This "clean expert + noise-robust learner" design outperforms using a single frozen VLM for both roles. LoRA-based finetuning results in balanced image editing, with a large improvement of **+4.42** HM scores compared to w/o LoRA.

(3) Stronger VLM backbone: To study the impact of different VLMs, we replaced LLaVA-7B with Qwen-VL-2.5-7B [1], later further improves alignment (CLIP-T: 27.60 → 27.91) and balanced editing (HM: 42.00 → 42.27), indicating that stronger semantic reasoning translates to better guidance.

(4) Context-preservation conditioning: Table 4(last two rows) present results with and without context-preservation conditioning to highlight its effectiveness. With conditioning, CS-P improves by +0.74, CLIP-T by +1.70, and HM by +1.81, demonstrating more balanced image editing.

Human Study: We have conducted a user study to understand the human-centric evaluations on the image editing task. The user study is conducted on the challenging RealEdit dataset with 30 subjects aged $20-45$. Evaluations are carried out across three dimensions: Instruction following, Non-instructed region preservation and Quality of generated edited image. Each participant selected the best-performing model in each evaluation metric case for the randomly provided 30 triplets of generated edited image, original image and the editing instruction. The triplets are provided for multiple instruction-based models such as InstructPix2Pix [2], MagicBrush [34], SuperEdit [19] and the proposed IABEdit.

Figure 6 shows the user responses for the best-performing model on 30 randomly selected images across all four instruction-based models. IABEdit achieves the highest Instruction Following rate of **50.00%**, demonstrating more accurate adherence to editing instructions than the competing methods. It also best preserves the non-instructed regions, achieving the highest preservation rate of **29.83%**. In terms of image quality, IABEdit attains **37.58%**, outperforming SuperEdit (**31.45%**). The slight variation in performance as compared to the GPT-4o model is observed due to the variation in sample size (i.e., 30 for human evaluation and 560 for GPT-4o evaluation) used for both evaluation methods. Human study results are in the supplementary.

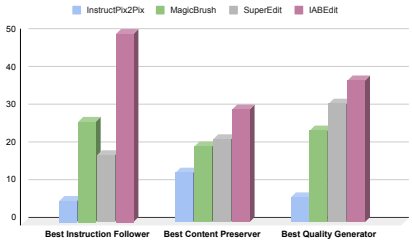

Fig. 6: User study results on RealEdit dataset. The values indicate the best performing model in randomly selected 30 images. Three metrics for human evaluations: Instruction Following, Non-Editing Content Preserving and Quality of Generated Edited image.

Training cost and Limitations: The limitations and training time comparisons are mentioned in the supplementary material.

5 Conclusion

IABEdit reframes instruction-guided editing as a problem of semantic alignment rather than conditioning. The argument is simple, a generative editor should not be trusted to follow an instruction it was never asked to semantically verify. By distilling spatially-aware descriptors from a frozen vision-language model and turning the residual into a gradient, IABEdit makes instruction satisfaction part of the training objective itself. Adherence is no longer assumed, it is measured, and minimized. Because this signal acts only during training, it leaves inference untouched and transfers across architectures, strengthening U-Net and MMDiT backbones alike. It even surpasses a proprietary baseline on the hardest case we tested, identity preservation under heavy occlusion, where curated benchmarks give way to real-world conditions. More broadly, IABEdit shows that gradient-aligned VLM distillation is a general way to supervise generation against meaning, not just pixels. The same principle extends naturally to video editing, 3D generation, and multimodal content creation, wherever verifying what was produced matters as much as producing it.

Acknowledgment

This research was supported by the FAE team, AMD India, through GPU resources and by the IndiaAI Mission and Meta through the Srijan: Centre of Excellence for Generative AI. Chiranjeev is partially supported through the PMRF Fellowship.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025)
2. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9630–9640 (2021)
4. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
5. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. ICLR (2023)
6. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4685–4694 (2019)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021)
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML. pp. 12606–12633 (2024)
10. Fu, T.J., Hu, W., Du, X., Wang, W., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. In: ICLR. pp. 54820–54833 (2024)
11. Geng, Z., Yang, B., Hang, T., Li, C., Gu, S., Zhang, T., Bao, J., Zhang, Z., Li, H., Hu, H., Chen, D., Guo, B.: Instructdiffusion: A generalist modeling interface for vision tasks. In: CVPR. pp. 12709–12720 (2024)
12. Gu, X., Li, M., Zhang, L., Chen, F., Wen, L., Luo, T., Zhu, S.: Multi-reward as condition for instruction-based image editing. In: ICLR. pp. 82243–82262 (2025)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
15. Huang, Y., Xie, L., Wang, X., Yuan, Z., Cun, X., Ge, Y., Zhou, J., Dong, C., Huang, R., Zhang, R., Shan, Y.: Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In: CVPR. pp. 8362–8371 (2024)
16. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Xie, C., Zhou, Y.: Hq-edit: A high-quality dataset for instruction-based image editing. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. pp. 73407–73424 (2025)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

18. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
19. Li, M., Gu, X., Chen, F., Xing, X., Wen, L., Chen, C., Zhu, S.: Superedit: Rectifying and facilitating supervision for instruction-based image editing. In: ICCV. pp. 19206–19215 (2025)
20. Li, M., Yang, T., Kuang, H., Wu, J., Wang, Z., Xiao, X., Chen, C.: Controlnet++: Improving conditional controls with efficient consistency feedback. In: ECCV. pp. 129–147 (2024)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
22. Manchanda, S., Bhagwatkar, K., Balutia, K., Agarwal, S., Chaudhary, J., Dosi, M., Chiranjeev, C., Vatsa, M., Singh, R.: D-lord: Dysl-ai database for low-resolution disguised face recognition. T-BIOM **6**(2), 147–157 (2024)
23. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML. pp. 16784–16804 (2022)
24. Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-g: Generating images in context with multimodal large language models. In: ICLR. pp. 40959–40974 (2024)
25. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR. pp. 1862–1874 (2024)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
27. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
28. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831 (2021)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (2022)
30. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS **35**, 36479–36494 (2022)
31. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: CVPR. pp. 8871–8879 (2024)
32. Singh, J., Zhang, J., Liu, Q., Smith, C., Lin, Z., Zheng, L.: Smartmask: Context aware high-fidelity mask generation for fine-grained object insertion and layout control. In: CVPR. pp. 6497–6506 (2024)
33. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: CVPR. pp. 22428–22437 (2023)
34. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. NeurIPS **36**, 31428–31449 (2023)

35. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., Xiong, C., Xu, R.: Hive: Harnessing human feedback for instructional visual editing. In: CVPR. pp. 9026–9036 (2024)
36. Zhao, H., Ma, X., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. NeurIPS **37**, 3058–3093 (2024)