DirectDrag: High-Fidelity, Mask-Free, Prompt-Free Drag-based Image Editing via Readout-Guided Feature Alignment

Anonymous WACV Applications Track submission

Paper ID 257

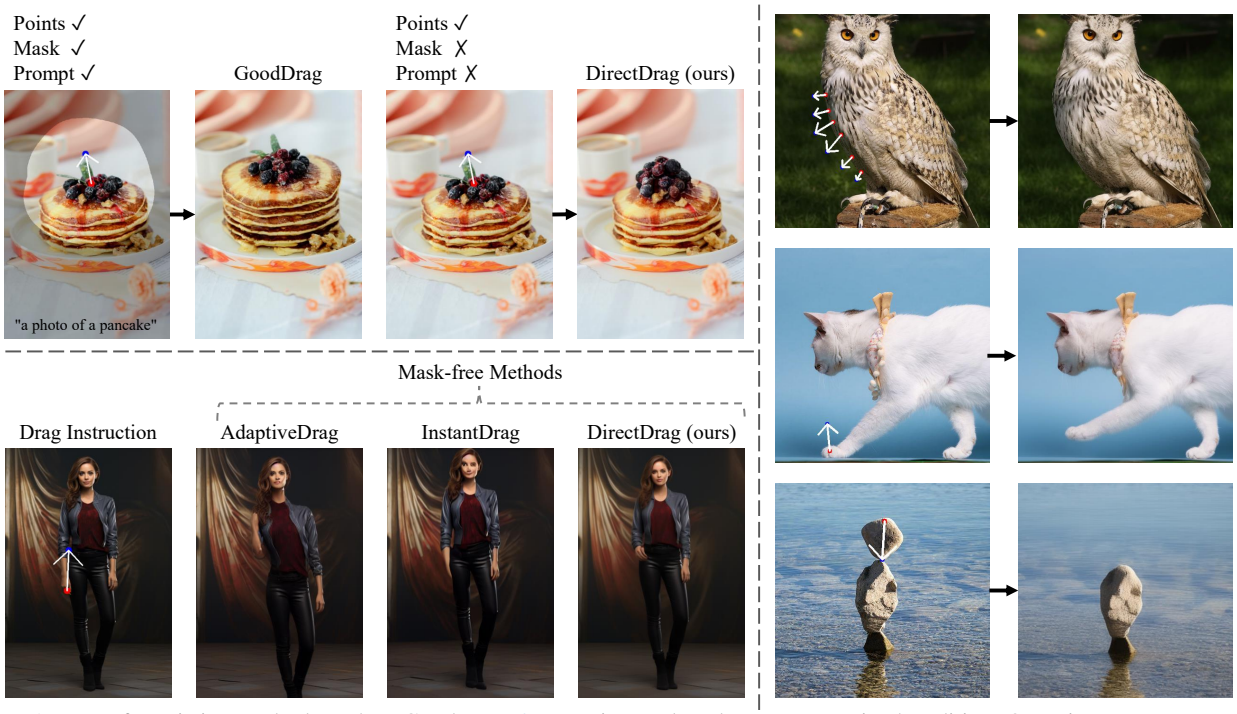

Figure 1. Up-Left: Existing methods such as GoodDrag [37] require mask and prompt to assist the editing. Our DirectDrag removes the dependency on mask and prompt, enabling more flexible editing while maintaining precise control. Bottom-left: Comparison with other mask-free methods, our method achieves more faithful and robust editing effects. Right: Additional qualitative results by DirectDrag.

Abstract

Drag-based image editing using generative models provides intuitive control over image structures. However, existing methods rely heavily on manually provided masks and textual prompts to preserve semantic fidelity and motion precision. Removing these constraints creates a fundamental trade-off: visual artifacts without masks and poor spatial control without prompts. To address these limitations, we propose DirectDrag, a novel mask- and prompt-free editing framework. DirectDrag enables precise and efficient manipulation with minimal user input while maintaining high image fidelity and accurate point alignment. DirectDrag introduces two key innovations. First, we design an Auto Soft Mask Generation module that intelligently infers editable regions from point displacement, automatically localizing deformation along movement paths while preserving contextual integrity through the generative model's inherent capacity. Second, we develop a Readout-Guided Feature Alignment mechanism that leverages intermediate diffusion activations to maintain structural consistency during point-based edits, substantially improving visual fidelity. Despite operating without manual mask or prompt, DirectDrag achieves superior image quality compared to existing methods while maintaining competitive drag accuracy. Extensive experiments on DragBench and real-world scenarios demonstrate the effectiveness and practicality of DirectDrag for high-quality, interactive image manipulation.

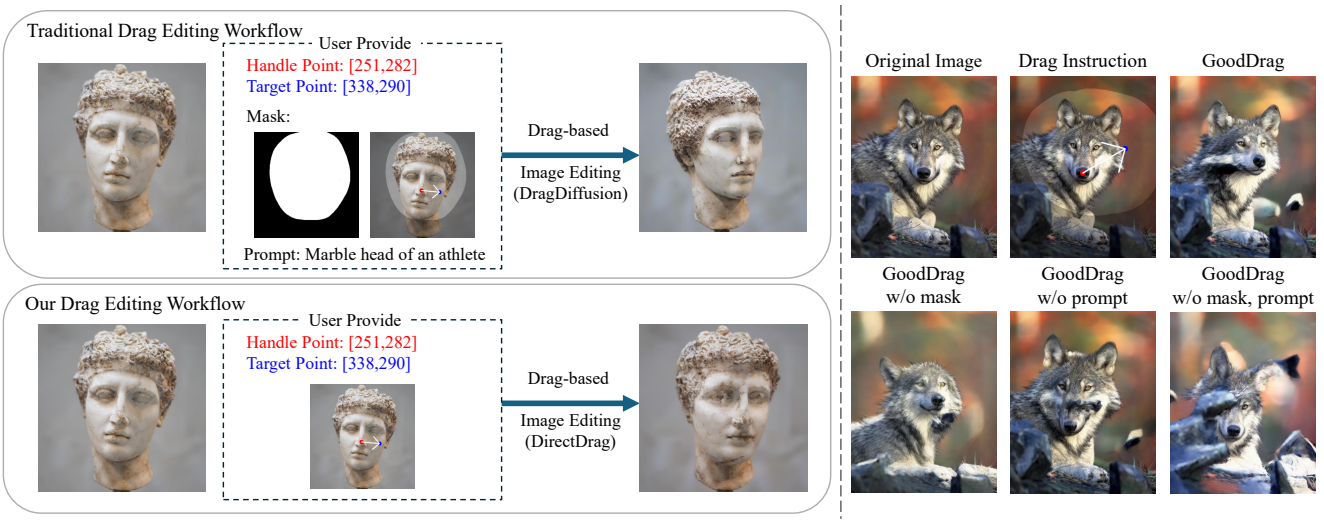

Figure 2. **Workflow Comparison. Left:** Traditional methods (e.g., DragDiffusion [29], GoodDrag [37]) rely on masks and prompts, increasing user burden. Our method simplifies the process by requiring only point inputs. **Right:** Removing masks leads to distortion, while omitting prompts reduces accuracy. We demonstrate these effects on GoodDrag [37] and also show the case without both inputs.

1. Introduction

Drag-based image editing has become a powerful and intuitive way to manipulate visual content. With recent advances in diffusion-based generative models [7, 25], this type of interaction has become increasingly precise and accessible. Unlike traditional text-to-image (T2I) methods [19, 23, 27], which rely on language to describe visual intentions, drag-based approaches provide direct and fine-grained control by allowing users to move a point from a source location to a desired target [18, 21, 29]. This enables a wide range of image modifications, including facial expression editing, object repositioning, content resizing, restoration, and data augmentation. Many existing methods still require users to provide additional information, such as an editable region mask and a text prompt, to ensure accurate and semantically coherent results [11, 12, 20, 33]. These extra inputs, while helpful in guiding the editing process, create two major sources of annotation overhead and instability. First, manually drawing an appropriate mask becomes particularly difficult when users want to edit multiple parts of an image at once. In such cases, designing a precise mask is not only time-consuming but also prone to errors. Poorly drawn masks often result in unexpected distortions or artifacts. Second, cues are often difficult to formulate accurately, especially when images contain multiple semantically rich regions. Describing a complex visual environment in one sentence is extremely challenging, and even slight errors in the cues may mislead the diffusion model and lead to poor results. In some scenarios—such as medical imaging or technical illustrations—there may not even be suitable natural language to express the intended change, making prompt-based control impossible. We find that removing the mask leads to noticeable loss of image fidelity (IF), while omitting the prompt significantly reduces point movement accuracy, reflected by increased mean distance (MD) scores. Therefore, eliminating these inputs, while desirable for simplifying user interaction, introduces real technical challenges. We illustrate these effects in Figure 2, where removing either the mask or the prompt leads to degraded visual quality or inaccurate drag results on a representative baseline (GoodDrag [37]). To address these issues, we present **DirectDrag**, a novel drag-based editing framework that operates in a fully mask-free and prompt-free setting. Our method maintains high visual quality and competitive spatial precision, all while requiring only minimal and intuitive input: handle and target points.

To achieve this, DirectDrag integrates three core technical components:

- An **Auto Soft Mask Generation** module that automatically infers editable regions based on point displacement. Rather than asking users to paint a mask manually, we localize deformation only along the path of movement, enhancing control where it matters most while relying on the generative model's capacity to preserve context elsewhere.
- A lightweight **Readout-Guided Feature Alignment** module that extracts intermediate diffusion features and aligns them based on spatial correspondence. This mechanism replaces the semantic guidance usually provided by prompt, helping the model maintain visual consistency and structure during editing.
- A **Latent Warpage Function**, adapted from prior work, which improves convergence and drag precision by initial-

izing latent codes with a geometry-aware deformation. This component offers a prompt-free alternative to guide the optimization process toward semantically plausible outcomes.

Together, these components allow DirectDrag to simplify the editing pipeline significantly. By removing the need for mask and prompt, we reduce the annotation burden and the risk of unstable or incorrect edits. As illustrated in Figure 1, our method outperforms existing mask-free approaches by producing more faithful and robust edits, even with minimal inputs. Despite having fewer user-provided signals, our approach achieves higher image fidelity than strong baseline. Although there is a slight trade-off in drag accuracy compared to full-input systems, the difference remains small. This suggests that our framework provides a favorable balance between usability and performance. We validate the effectiveness of DirectDrag through extensive experiments on DragBench and real-world images, confirming its potential for practical and scalable interactive editing.

2. Related Work

2.1. Generative Image Models and Image Editing

Generative image models, particularly GANs and diffusion models, have significantly enhanced image synthesis and editing capabilities. GANs [6, 10] provide fast generation, but stable reversible editing is often difficult to achieve. Diffusion models [7, 25, 27] show outstanding fidelity through iterative denoising of latent codes. These models form the basis of interactive image editing applications. Image editing techniques can be divided into content-aware and content-free methods: Content-Aware Editing includes object manipulation, spatial transformation, inpainting, and style transfer. Text-prompted editing methods (e.g., InstructPix2Pix [1, 27]) and user-guided approaches fall into this category. Content-Free Editing focuses on customization using user-specified images or attributes. Examples include subject-driven personalization (e.g., DreamBooth [26]) and attribute-driven fine-tuning.

2.2. Drag-based Image Editing

Drag-based image editing methods enable users to control image structures by dragging specific points to target locations. DragGAN [21] first proposed a latent code optimization framework with point tracking based on GANs, but struggled with generalizing to real-world inputs. DragDiffusion [29] and DragonDiffusion [18] extended this paradigm to diffusion models, improving structural manipulation and semantic controllability through prompt conditioning and denoising-based alignment. Subsequent methods have aimed at improving editing quality and robustness. DragNoise [13] reduces computational cost by optimizing bottleneck features of the U-Net instead of full latents. GoodDrag [37] alternates between dragging and denoising to prevent error accumulation and preserve image fidelity. GDrag [11] follows a training-free approach that tackles intention and content ambiguity through atomic manipulation taxonomy and dense trajectory estimation.

Other works focus on enhancing editing efficiency. DiffEditor [17] reduces optimization time by decreasing the number of diffusion steps. FastDrag [38] uses a one-step feed-forward generation approach for instant edits. LightningDrag [28] treats editing as conditional generation trained on large-scale video data for fast, accurate results. EEdit [34] accelerates editing by reducing spatial and temporal redundancy through region caching and inversion step skipping.

2.3. Mask-Free Drag-Based Image Editing

Recent works have proposed removing manually provided masks to simplify the drag editing pipeline while preserving semantic and structural control. EasyDrag[8] focuses on user-friendliness by eliminating the need for masks and tuning procedures such as LoRA[9]. It leverages pretrained diffusion models without architectural modifications and achieves better editing precision and visual quality than DragDiffusion [29]. However, it still requires a text prompt to maintain semantic guidance, which limits usability in prompt-free scenarios. In addition, EasyDrag relies on ControlNet [35], which introduces considerable memory overhead during inference.

InstantDrag [30] improves editing speed by introducing an optimization-free pipeline that takes only an image and a drag instruction as input. It uses a drag-conditioned optical flow network followed by a flow-guided diffusion model to achieve fast and realistic edits. While it avoids mask and prompt, InstantDrag must retrain a dedicated diffusion model on large-scale video data, significantly increasing parameter count and training cost. Moreover, it often requires multiple drag instructions to produce stable results, reducing its effectiveness in sparse user-interaction settings. AdaptiveDrag [2] introduces automatic mask generation using superpixel segmentation by SAM2 [24] and incorporates semantic-aware latent optimization guided by adaptive steps and a specialized loss. Although it improves localization accuracy and generalization across categories, AdaptiveDrag depends on external segmentation models and still requires textual prompt for semantic alignment, resulting in additional computational overhead.

While these methods effectively reduce the need for manual mask input, they either rely on prompt, introduce heavy architectural modifications, or require extra modules such as segmentation or flow estimation. In contrast, **DirectDrag** adopts a lightweight and fully mask-free and prompt-free framework that maintains high image fidelity and competitive drag precision. It achieves this through automatic soft mask generation, readout-guided feature align-

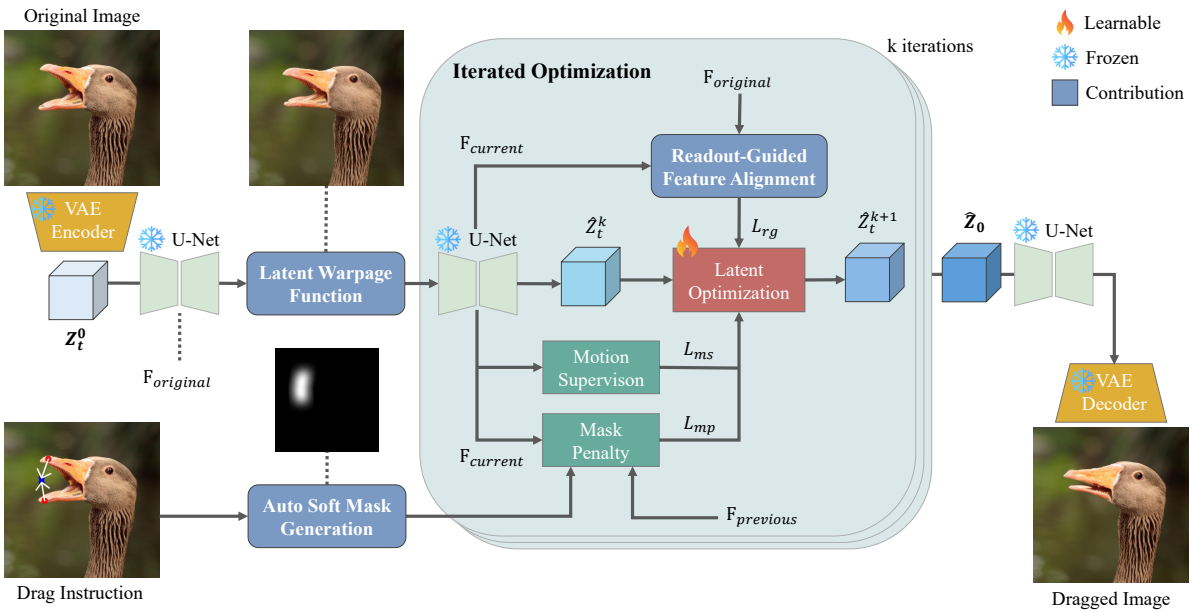

Figure 3. **Overview of the proposed DirectDrag framework**. Given an input image and point pairs, we apply DDIM inversion to obtain latent codes, initialize editing via latent warpage function and generate soft mask, then iteratively apply drag and denoising guided by motion supervision and feature alignment.

ment, and latent warpge function introducing only a minimal auxiliary module, far more efficient and compact than the large-scale components used in existing approaches.

3. Method

3.1. Overview

We propose **DirectDrag**, a fully mask-free and prompt-free framework for drag-based image editing. Unlike previous diffusion-based methods [4, 18, 29, 37], which rely on hand-crafted mask or prompt, our method simplifies the pipeline while preserving editing quality. As shown in Figure 3, the process begins by applying DDIM inversion [31] to encode the input image into latent space. A geometry-aware latent warpage function (LWF) initializes the latent code, and an auto soft mask generation module estimates the editable region based on point displacement—removing the need for manual masks. We adopt the AIDD strategy [37] (Alternating Inversion and Drag-Denoise) to optimize the latent representation iteratively. During each step, drag loss encourages point movement, while our readout-guided Feature alignment module extracts intermediate diffusion features to maintain visual consistency. These components work together to preserve fidelity and precision even without prompts or segmentation inputs.

Compared to prior work that introduces architectural changes [30] or external segmentation tools [2], DirectDrag remains lightweight and modular, while achieving strong fidelity and alignment performance across diverse examples.

3.2. Latent Diffusion and DDIM Inversion

Denoising Diffusion Probabilistic Models (DDPMs) [7] have demonstrated strong generative capabilities by modeling the image generation process as a gradual denoising of random noise. However, operating directly in pixel space is computationally expensive. To improve efficiency, Latent Diffusion Models (LDMs) [25] encode the image $\mathbf{x}_0$ into a lower-dimensional latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ using a pretrained VAE encoder $\mathcal{E}$. The diffusion process is then carried out in the latent space as a Markov chain over T timesteps, where the marginal likelihood is expressed as:

$$p_\theta(\mathbf{z}_0) = \int p_\theta(\mathbf{z}_{1:T})\, d\mathbf{z}_{1:T}, \tag{1}$$

where each latent variable $\mathbf{z}_t$ is obtained by progressively adding Gaussian noise to $\mathbf{z}_0$ using a forward process defined as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t$ denotes the cumulative product of noise schedule coefficients up to timestep t.

To enable editing from real images, we adopt deterministic DDIM inversion [31], which reverses the diffusion process to recover latent trajectories. This allows us to initialize the editing process from a clean latent code $\mathbf{z}_0$ without requiring random sampling. Since our method does not rely on prompts, DDIM inversion is performed in a

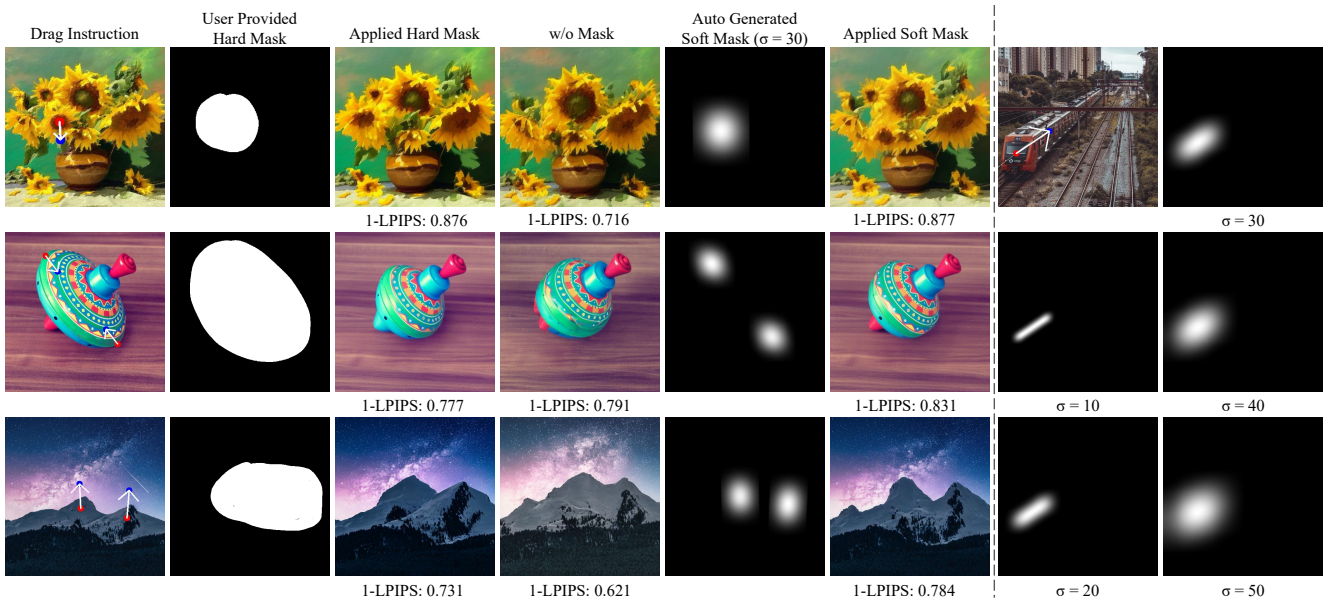

Figure 4. **Effect of our Soft Mask.** Left: Compared to no masking and user provide hard mask, applying the generated soft mask significantly improves visual fidelity and structure preservation, as reflected by higher image fidelity scores (1-LPIPS↑). Right: Visualization of soft masks under different drag configurations and Gaussian widths (σ), illustrating their adaptiveness to motion magnitude and direction.

prompt-free setting, enabling faithful reconstructions and providing a robust starting point for subsequent drag-based manipulation.

3.3. Drag-based Image Editing

Our method builds upon prior drag-based diffusion editing approaches [29, 37], where user-specified handle points are iteratively moved toward target locations by optimizing latent features in the diffusion model. To guide this deformation process, we incorporate three key components: motion supervision, alternating inference-driven denoising (AIDD), and feature-based point tracking.

Motion Supervision. We adopt a multi-step motion supervision loss to encourage the features at displaced handle points to match those at their original locations. This supervision helps align internal features with the intended motion trajectory:

$$\mathcal{L}_{\text{ms}} = \sum_{i=1}^{n} \sum_{q} \left\| \mathcal{F}_{q+d_i}(\hat{\mathbf{z}}_t^k, \hat{\mathbf{c}}^k) - \text{sg}(\mathcal{F}_q(\hat{\mathbf{z}}_t^k, \hat{\mathbf{c}}^k)) \right\|_1, \quad (3)$$

where $\mathcal{F}_q$ denotes the U-Net features extracted at location q, and d_i is the displacement vector of the i-th handle point.

AIDD Optimization Schedule. To prevent noise accumulation and preserve global image structure, we adopt the Alternating Inference-Driven Denoising (AIDD) schedule proposed in GoodDrag [37]. Rather than performing continuous updates in the latent space, AIDD interleaves B drag steps with periodic denoising steps. This scheduling helps retain proximity to the image manifold and stabilizes optimization. At each drag step, we apply a patch-level alignment loss:

$$\mathcal{L}_{\text{drag}} = \sum_{i} \left\| \mathcal{F}_{\Omega(\mathbf{p}_i + \delta\mathbf{p}_i)} - \text{sg}(\mathcal{F}_{\Omega(\mathbf{p}_i)}) \right\|_1, \quad (4)$$

where $\Omega(\cdot, r_1)$ extracts a spatial patch of radius r_1, and $\delta\mathbf{p}_i^k$ is the displacement from the initial handle position to its target.

Point Tracking. We also incorporate the point tracking mechanism from GoodDrag [37] to maintain semantic consistency throughout the editing trajectory. Instead of keeping handle points fixed across iterations, we dynamically update each point's position by matching its initial diffusion features with features from nearby locations in the current timestep. This allows the model to follow the semantic content even as the image structure evolves during optimization. The detailed formulation of this tracking algorithm is provided in the supplementary material.

Together, motion supervision, AIDD scheduling, and feature-based tracking form the core optimization loop that enables precise point-based editing while preserving image quality and structural coherence.

3.4. Auto Soft Mask Generation

In drag-based editing, prior methods often rely on user-provided hard mask to confine deformation. However, even with these mask, diffusion models tend to produce unintended changes in unrelated regions due to weak spatial

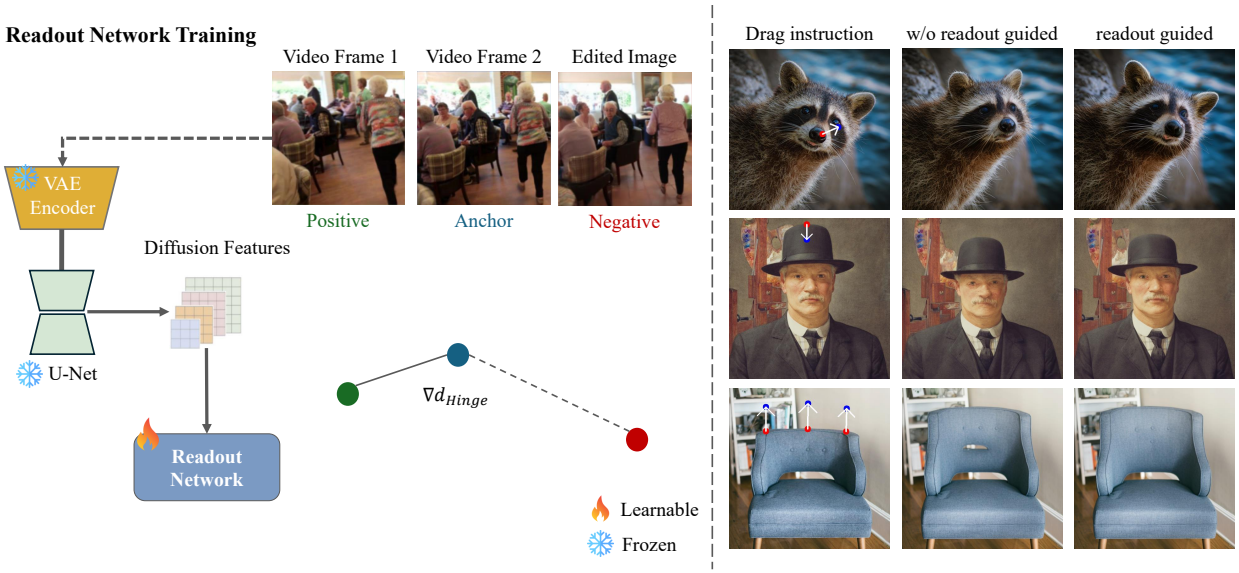

Figure 5. **Readout Network Training and Effect.** Left: We train the readout network using a triplet loss on diffusion features extracted from video frames (anchor, positive) and edited images (negative). Right: Incorporating readout guidance preserves appearance details and improves structural consistency during dragging.

constraints. In practice, omitting mask altogether leads to even more severe artifacts, such as missing objects, hallucinated structures, or drastic changes in color and composition—as shown in Fig. 4.

To improve usability while reducing over-editing, we propose to generate a soft spatial mask $M \in [0,1]^{H\times W}$ directly from the drag instructions. This removes the burden of manual annotations and ensures localized structural control. Specifically, for each handle–target pair $(\mathbf{h}_i, \mathbf{t}_i)$ with coordinates (x_0, y_0) and (x_1, y_1), we interpolate $N = \max(|x_1 - x_0|, |y_1 - y_0|) + 1$ points along the linear path connecting them:

$$\tilde{M}(x_k, y_k) = 1, \quad \text{where}$$
$$(x_k, y_k) = \lfloor (1-\alpha_k)(x_0, y_0) + \alpha_k (x_1, y_1) \rceil, \tag{5}$$
$$\alpha_k = \frac{k}{N-1} = \frac{k}{\max(|x_1 - x_0|, |y_1 - y_0|)}. \tag{6}$$

We accumulate $\tilde{M}$ from all point pairs, then apply a Gaussian filter followed by normalization to form the final soft mask M:

$$M = \frac{\text{GaussianBlur}(\tilde{M}, \sigma)}{\max\left(\text{GaussianBlur}(\tilde{M}, \sigma)\right)}. \tag{7}$$

The resulting soft mask softly highlights the regions along dragging trajectories, enforcing smooth, localized constraints without introducing sharp editing boundaries. While this design significantly reduces unintended edits, it has its limitations: the linear interpolation path may not fully cover the deformable object, especially for complex geometries. Nevertheless, we argue that the primary role of a mask is to localize major structural changes—not to precisely capture every affected pixel. In fact, over-constraining the optimization via strict loss masking can conflict with the global nature of latent updates in diffusion models, sometimes degrading drag precision instead of improving it. Our lightweight mask acts as a guiding prior, with finer control delegated to subsequent alignment mechanisms.

3.5. Readout-Guided Feature Alignment

Although the soft mask improves visual fidelity and local stability, it often fails to suppress subtle background artifacts or hallucinated textures, as illustrated in Fig. 5. To address this, we incorporate a feature alignment mechanism based on Diffusion Hyperfeatures [15] and Readout Guidance [14].

Readout Network. Following Luo *et al.* [14], we use a lightweight readout network trained to extract appearance-preserving features from intermediate U-Net layers of a frozen denoiser. Supervision is provided via a triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max\big(0,\, D(F(I_a), F(I_p)) - D(F(I_a), F(I_n)) + \delta\big) \tag{8}$$

where $F(\cdot)$ is the readout head output, D is cosine distance, and I_p, I_n are positive and negative samples. Negative examples are generated by SDEdit [16], which perturbs appearance while preserving structure. Readout Guidance [14] use training data from the Pascal VOC dataset [5].

Method	Venue	Mask	Prompt	IF↑	CLIP SIM↑	MD↓	Model Params	Tuning Params
DragDiffusion [29]	CVPR'24	✓	✓	0.883	0.977	32.87	865M	0.07M
FreeDrag [12]	CVPR'24	✓	✓	0.897	0.977	33.82	865M	0.07M
DiffEditor [17]	CVPR'24	✓	✓	0.877	0.966	31.70	865M	0.07M
DragNoise [13]	CVPR'24	✓	✓	0.899	0.972	37.92	865M	0.33M
FastDrag [38]	NeurIPS'24	✓	✓	0.859	0.963	32.66	865M	0
GoodDrag [37]	ICLR'25	✓	✓	0.869	0.977	25.28	865M	0.07M
DragText [3]	WACV'25	✓	✓	0.870	0.971	34.25	865M	0.12M
LightningDrag [28]	ICML'25	✓	✓	0.881	0.970	29.95	933M	933M
Mask-free methods								
EasyDrag* [8]	CVPR'24	✗	✓	0.882	–	34.44	1770M	**0.07M**
AdaptiveDrag [2]	ArXiv'24	✗	✓	0.867	0.975	33.94	1168M	**0.07M**
InstantDrag [30]	SIGGRAPH Asia'24	✗	✗	0.878	0.968	30.41	914M	914M
DirectDrag (ours)$_{\text{w/o LWF}}$	–	✗	✓	**0.918**	**0.982**	31.91	**871M**	5.97M
DirectDrag (ours)	–	✗	✗	0.891	0.976	**29.65**	**871M**	5.97M

Table 1. **Quantitative evaluation** on the DragBench [29] dataset. IF = 1 - LPIPS. CLIP SIM = CLIP [22] Similarity. MD = Mean Distance. ✓: Required, ✗: Not Required. LWF: Latent Warpage Function. Model Params: Total parameters used in model. Tunning Params: Parameters require to training in correspond method. * means scores are taken from the another publication.

Inference-Time Guidance. During editing, we extract intermediate features from the original image $\mathbf{z}_t^0$ (before any dragging) and use them as the reference for appearance alignment. For each optimization step, the current latent $\bar{\mathbf{z}}_t$ is passed through the readout network, and the following loss is applied:

$$\mathcal{L}_{\text{rg}} = \|F(\bar{\mathbf{z}}_t^{\text{k}}) - F(\mathbf{z}_t^0)\|_2^2, \tag{9}$$

where $F(\cdot)$ denotes the readout network's output from selected U-Net layers (e.g., down3 to up2). This encourages the edited latent to stay visually close to the original appearance, mitigating hallucination and identity drift. Unlike Readout Guidance [14], which is designed for one-shot diffusion and prone to hallucinations, our approach integrates readout features into a multi-step optimization framework. This allows better convergence and reduces artifacts, especially in challenging scenes. The guidance is effective without modifying the diffusion backbone, introducing only minor overhead while improving appearance stability.

3.6. Latent Warpage Function

To initialize the latent with geometry-aware deformation, we adopt the latent warpage function (LWF) from FastDrag [38]. For each masked pixel p_j in latent space, its displacement $\mathbf{v}_j$ is computed as a weighted combination of drag vectors $\mathbf{d}_i = e_i - s_i$:

$$\mathbf{v}_j = \sum_{i=1}^{k} w_j^i \cdot \lambda_j^i \cdot \mathbf{d}_i, \tag{10}$$

where w_j^i is the inverse distance weight to handle s_i, and λ_j^i is a stretch factor based on geometric intersections.

Unlike the original latent warpage function, which often over-applies displacement and harms fidelity, we scale the drag vector with a ratio ρ:

$$\mathbf{d}_i' = \rho \cdot (e_i - s_i), \tag{11}$$

producing a gentler shift in latent space. This mitigates early semantic drift and improves convergence. Empirically, this initialization reduces mean distance error and enables more stable drag optimization in subsequent steps.

4. Experiments

4.1. Implementation Details

We build on Stable Diffusion v1.5 [25] and run all experiments on an NVIDIA RTX 4090. Our pipeline follows DDIM inversion with 50 inference steps and guidance scale 1.0. We highlight three key settings: (1) Soft Mask: Gaussian blur with $\sigma = 30$ ensures smooth region transitions. (2) Readout-Guided Weight: The readout guidance loss is scaled by 350 before adding to the main objective. (3) Latent Warpage Function: To reduce over-drag during initialization, we apply 15% of the displacement vector from handle to target. All other parameters follow settings from baseline (GoodDrag [37]).

4.2. Quantitative Evaluation

We evaluate on DragBench [29] using (1) 1-LPIPS [36] for perceptual similarity, (2) CLIP [22] Similarity for

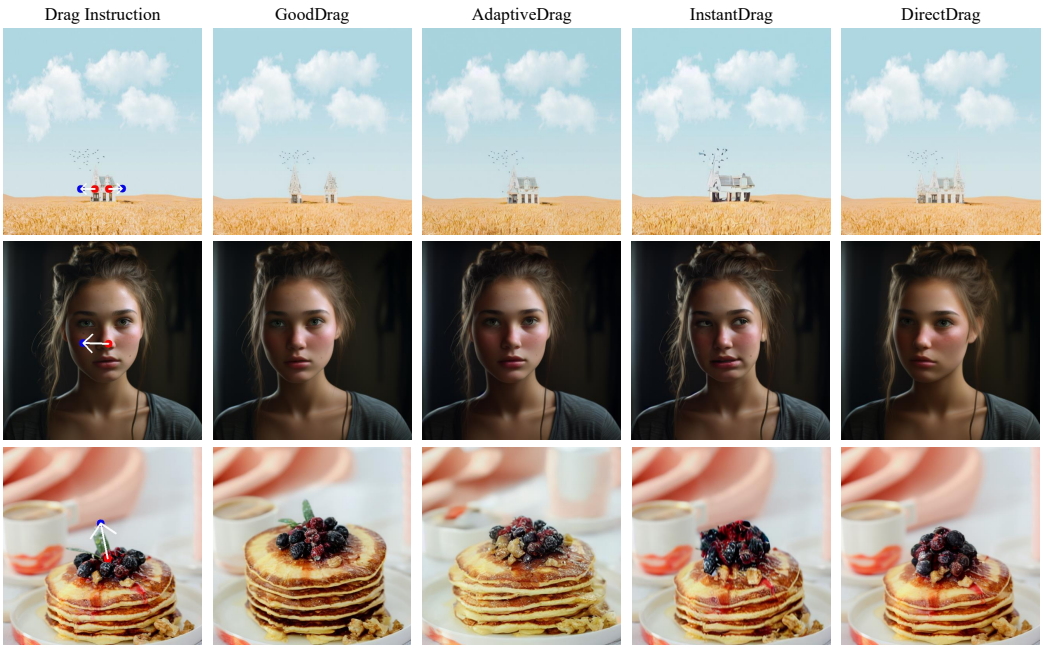

Figure 6. **Qualitative comparison**. Compared to the baseline (*GoodDrag* [37]) and mask-free methods (*AdaptiveDrag* [2], *InstantDrag* [30]), our method *DirectDrag*)

Method	SM	RG	LWF	IF↑	CLIP SIM↑	MD↓
Baseline				0.789	0.963	24.74
+ Soft Mask	✓			0.895	0.979	31.35
+ Readout Guide	✓	✓		**0.918**	**0.982**	33.75
+ Readout Guide $_{+prompt}$	✓	✓		**0.918**	**0.982**	31.91
+ Latent Warpage	✓	✓	✓	0.891	0.976	29.65
+ Latent Warpage $_{+prompt}$	✓	✓	✓	0.891	0.975	**29.18**

Table 2. **Ablation study of *DirectDrag***. Baseline indicates GoodDrag [37] without mask and prompt.

semantic consistency, and (3) MD [21] for dragging accuracy using DIFT [32]. As shown in Table 1, DirectDrag perfrom **state-of-the-art** result in mask-free methods. Despite working in minimal input conditions, DirectDrag matches or exceeds mask-based and prompt-based methods in image fidelity and drag accuracy.

4.3. Qualitative Results

Fig. 6 compares DirectDrag to GoodDrag [37] (baseline with mask and prompt) and two mask-free methods, AdaptiveDrag [2] and InstantDrag [30]. While the latter often suffers from distortions or incomplete motion, our method achieves more accurate and stable edits. Across diverse cases—motion, face, and object deformation—*DirectDrag* maintains background consistency and visual detail, confirming its advantage in prompt- and mask-free editing.

4.4. Ablation Study

Table 2 shows the impact of each component in *DirectDrag*. The soft mask significantly improves visual fidelity, while readout guidance helps preserve appearance but slightly reduces motion accuracy. Latent warpage function improves spatial precision with minimal degradation in image quality. We also tested a variant using prompt conditioning, showing that our latent warpage function can effectively replace prompt for improving drag accuracy. Overall, our final setup offers the best trade-off between fidelity and accuracy in a fully mask-free and prompt-free setting.

5. Conclusion

We presented *DirectDrag*, a lightweight framework for drag-based image editing that operates without manual mass or prompt. By integrating automatic soft mask generation, readout-guided feature alignment, and a latent warpage function, our method achieves high visual fidelity and competitive dragging accuracy. Extensive experiments demonstrate that *DirectDrag* provides a practical and effective solution for intuitive image manipulation, balancing usability, precision, and quality.

Limitation. In some cases, our method may over-preserve visual fidelity, resulting in insufficient deformation. Additionally, strong geometric warping can occasionally cause texture detail loss.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 3

[2] Yining Chen, Qi Wang, Hao Zhu, Hongxu Lin, and Yibing Xu. Adaptivedrag: Mask-free point-based image editing with editable region localization. *arXiv preprint arXiv:2410.12696*, 2024. 3, 4, 7, 8

[3] Gayoon Choi, Taejin Jeong, Sujung Hong, and Seong Jae Hwang. Dragtext: Rethinking text embedding in point-based image editing. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 441–450. IEEE, 2025. 7

[4] Yutao Cui, Xiaotong Zhao, Guozhen Zhang, Shengming Cao, Kai Ma, and Limin Wang. Stabledrag: Stable dragging for point-based image editing. In *European Conference on Computer Vision*, pages 340–356. Springer, 2024. 4

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge 2012 (voc2012) results (2012), 2011. 6

[6] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 4

[8] Xingzhong Hou, Boxiao Liu, Yi Zhang, Jihao Liu, Yu Liu, and Haihang You. Easydrag: Efficient point-based manipulation on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8404–8413, 2024. 3, 7

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 3

[10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[11] Xiaojian Lin, Hanhui Li, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Gdrag: Towards general-purpose interactive editing with anti-ambiguity point diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3

[12] Pengyang Ling, Lin Chen, Pan Zhang, Huaian Chen, Yi Jin, and Jinjin Zheng. Freedrag: Feature dragging for reliable point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6860–6870, 2024. 2, 7

[13] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6743–6752, 2024. 3, 7

[14] Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8217–8227, 2024. 6, 7

[15] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023. 6

[16] 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. 6

[17] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8488–8497, 2024. 3, 7

[18] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *International Conference on Learning Representations*, 2024. 2, 3, 4

[19] Alex 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. 2

[20] Shen Nie, Hanzhong Allan Guo, Cheng Lu, Yuhao Zhou, Chenyu Zheng, and Chongxuan Li. The blessing of randomness: Sde beats ode in general diffusion-based image editing. In *International Conference on Learning Representations*, 2024. 2

[21] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2, 3, 8

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *International Conference on Machine Learning*, 2025. 3

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022. 2, 3, 4, 7

[26] 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*, pages 22500–22510, 2023. 3

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 3

[28] Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent YF Tan, and Jiashi Feng. Lightningdrag: Lightning fast and accurate drag-based image editing emerging from videos. In *International Conference on Machine Learning*, 2025. 3, 7

[29] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 2, 3, 4, 5, 7

[30] Joonghyuk Shin, Daehyeon Choi, and Jaesik Park. Instantdrag: Improving interactivity in drag-based image editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–10, 2024. 3, 4, 7, 8

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 4

[32] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 8

[33] Zixuan Wang, Duo Peng, Feng Chen, Yuwei Yang, and Yinjie Lei. Training-free dense-aligned diffusion guidance for modular conditional image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13135–13145, 2025. 2

[34] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*, 2025. 3

[35] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[37] Zewei Zhang, Huan Liu, Jun Chen, and Xiangyu Xu. Goodrag: Towards good practices for drag editing with diffusion models. In *International Conference on Learning Representations*, 2025. 1, 2, 3, 4, 5, 7, 8

[38] Xuanjia Zhao, Jian Guan, Congyi Fan, Dongli Xu, Youtian Lin, Haiwei Pan, and Pengming Feng. Fastdrag: Manipulate anything in one step. *Advances in Neural Information Processing Systems*, 37:74439–74460, 2024. 3, 7