ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889818

Textual and Visual Prompt Fusion for Image Editing via Step-Wise Alignment

Zhanbo Feng*, Zenan Ling†, Xinyu Lu*, Ci Gong†, Feng Zhou‡, Wugedele Bao§,
Jie Li*, Fan Yang*, and Robert C. Qiu†
*Department of Computer Science and Engineering, Shanghai Jiao Tong University
†EIC, Huazhong University of Science and Technology §School of Computer Science, Hohhot Minzu College

***Abstract*—The use of denoising diffusion models is becoming increasingly popular in the field of image editing. However, current approaches often rely on either image-guided methods, which provide a visual reference but lack control over semantic consistency, or text-guided methods, which ensure alignment with the text guidance but compromise visual quality. To resolve this issue, we propose a framework that integrates a fusion of generated visual references and text guidance into the semantic latent space of a *frozen* pre-trained diffusion model. Using only a tiny neural network, our framework provides control over diverse content and attributes, driven intuitively by the simple prompt. Compared to state-of-the-art methods, the framework generates images of higher quality while providing realistic editing effects across various benchmark datasets. The code is available at https://github.com/SadAngelF/Editing-via-Step-Wise-Alignment.**

***Index Terms*—Diffusion Model, Multi-modal, Image Editing, Generative Models, Zero Shot.**

I. Introduction

Manipulating real-world images with natural language has long been a challenge in image processing. Recently, denoising diffusion models (DDMs) have shown substantial success in text-to-image tasks, exemplified by models like Imagen [1], Dall-E [2], and Stable Diffusion [3]. These text-to-image models produce diverse, highly coherent, and realistic images that align well with text prompts. However, manipulating attributes on real images is still a significant challenging.

With the advancement of Natural Language Processing (NLP) techniques, e.g., GPT [4], [5], [6], considerable effort has been invested in text-guided image editing [7], [8]. Many previous works [9], [10], [11], [12] have developed image-editing techniques guided by textual prompts. However, they tend to neglect the importance of visual references. Despite maintaining semantic fidelity, text-guided methods struggle to learn fine-grained visual patterns from textual features in the absence of a visual prior. Textual semantics alone provide insufficient visual reference, leading to imprecise semantic manipulation. Text-guided editing is especially prone to failure when the target semantic is outside the domain.

On the other hand, image-guided editing can easily perform style transfer [13], [14], [15], [16], inpainting [17], and item replacement [18], [19]. With visual reference, generators insert ready-made visual patterns into images directly. Specifically, Taming Encoder [18] embeds specified elements into the target image by encoding a reference image. VISII [15] blends both textual and visual prompts to learn a style transfer from paired examples, representing the "before" and "after" images of an edit. However, image-guided approaches lack intuitive control over semantic consistency, making it ambiguous to specify which attribute should be referenced from the image.

In this paper, we propose Step-Wise Alignment (SWA) that integrates both visual references and text guidance into the semantic latent space of a *frozen* pre-trained diffusion model. Our method leverages text guidance to provide intuitive control over semantic consistency, while refining the alignment between the text features and the semantic latent space of the diffusion model by incorporating a visual reference. Our main contributions are as follows:

1) We introduce a framework that integrates a fusion of visual and textual prompts for attribute editing on real images.
2) We propose Step-Wise Alignment to align the text-image fusion features and the semantic latent space of the *frozen* diffusion model. Benefiting from zero-shot optimization, SWA avoids collecting the data of specific attributes.
3) SWA is evaluated on various benchmark datasets, including CelebA-HQ, LSUN-church, and LSUN-bedroom, and it outperforms state-of-the-art methods in terms of image quality and attribute manipulation.

II. Methodology

A. Problem Definition

Given an image $i_{\text{edit}} \in \mathbb{R}^{m\times n}$ and an attribute t_{attr}, our primary objective is to modify i_{edit} according to the attribute t_{attr}, resulting in an edited image, denoted as i_{out}.

Z. Ling is supported by the National Natural Science Foundation of China (via NSFC-62406119), the Natural Science Foundation of Hubei Province (2024AFB074), and the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

J. Li has been partially supported by NSFC Grants 61932014, 62232011 and the Shanghai Science and Technology Innovation Action Plan Grant 24BC3201200.

W. Bao is partially Supported by Natural Science Foundation of Inner Mongolia Autonomous Region of China Grant 2021MS06001.

R. C. Qiu would like to acknowledge the National Natural Science Foundation of China (via NSFC12141107), the Interdisciplinary Research Program of HUST (2023JCYJ012), the Key Research and Development Program of Guangxi (GuiKe-AB21196034). F. Zhou was supported by the National Natural Science Foundation of China Project (via NSFC-62106121) and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001). *(Corresponding authors: Zenan Ling , Jie Li)*

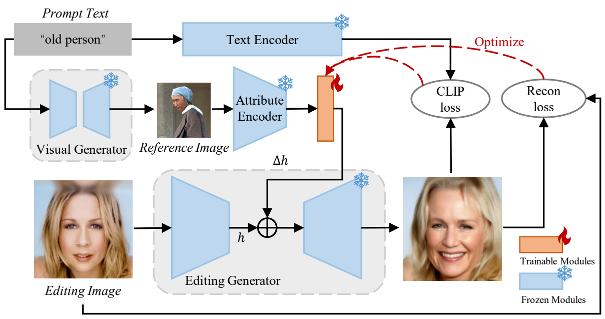


Fig. 1: **The framework of SWA.** The reference image is encoded into features Δh. Then, Δh are integrated into the latent features h of the editing image. The textual prompt contributes semantic information for the manipulation process.

Directly adding noise to i_{edit} and performing denoising within a *frozen* diffusion model is not a feasible approach for attribute editing. The reason is that such a *frozen* model may lack semantic relevance and may fail to retain the desired attributes. In addition, the added noise can distort the image and introduce undesirable artifacts. Therefore, we propose a framework to refine the alignment between text features and the semantic latent space of the diffusion model by incorporating a visual reference.

Therefore, we aim to optimize the reverse process of a *frozen* diffusion model for meaningful attribute editing. A typical reverse process in diffusion model is

$$\begin{aligned} x_{t-1} = \sqrt{\alpha_{t-1}/\alpha_t}\left(x_t - \sqrt{1-\alpha_t}\epsilon_\theta(x_t,t)\right) \\ + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(x_t,t) + \sigma_t\epsilon_t, \end{aligned} \tag{1}$$

where $\epsilon_t \sim \mathcal{N}(\mathbf{0},\boldsymbol{I})$ is standard Gaussian noise, α_t is the parameter based on the forward process, $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$, and $\epsilon_\theta(x_t,t)$ is a neural network to predict the noise in x_t. In the following, we introduce our framework to optimize the process above.

B. Framework

As illustrated in Figure 1, our framework consists of four key components: Text Encoder, Visual Generator, Attribute Encoder, and Editing Generator. To obtain the visual attributes corresponding to the designated attribute, we utilize a text-image model as the Visual Generator. Furthermore, when the attribute involves adding embellishments, such as the glasses, a reference image can be manually provided to ensure consistency of embellishments during editing. Both the textual prompt and the visual prompt, which are encoded by the Text Encoder and the Attribute Encoder respectively, are employed for editing in the Editing Generator.

Text Encoder: The text prompt t from the attribute t_{attr} is encoded into a vector for the purpose of calculating loss with the target image. In this paper, CLIP [20] is used as the Text Encoder.

In Text Encoder, let E_T be a text encoder with vocabulary V. The attribute t_{attr} is a sequence of phrases $t_{\text{attr}} = (s_1, \ldots, s_k)$ with $s_i \in V$; for example, $k = 1$ if the attribute t_{attr} is "glasses." Similar to the prompt in NLP [21], we define a *pattern* as a function P that takes t_{attr} as input and outputs two phrases or sentences $t_{\text{source}}, t_{\text{target}} = P(t_{\text{attr}}) \in V$ as the text prompt $t = (t_{\text{source}}, t_{\text{target}})$. For example, the pattern $P(t_{\text{attr}}) =$ ("a person", "a person with t_{attr}") will be used for the attribute t_{attr} of the person. Given an input attribute $t_{\text{attr}} =$ "glasses", then the text prompt will be $P(t_{\text{attr}}) =$ ("a person", "a person with glasses"). After that, $t_{\text{source}} =$ "a person" and $t_{\text{target}} =$ "a person with glasses". Both of them compose the text prompt t. Using the Text Encoder, the text prompt will be encoded as $E_T(t_{\text{source}}) \in \mathbb{R}^d$ and $E_T(t_{\text{target}}) \in \mathbb{R}^d$.

Visual Generator: To obtain visual features of the designated attribute, a text-image model is used as the Visual Generator. Large generative models are known for strong robustness and generalization in conditional generation [1], [3]. Despite their limitations in accurately detecting or modifying attributes, they can effectively generate the corresponding visual features. In this paper, we use UniDiffuser [9] as the Visual Generator, which generates images by one model, benefiting from the marginal, conditional, and joint distributions determined by multi-modal data. The reference image i_{ref} is sampled from the conditional distribution $p(x_0|t_{\text{target}})$ and denoised by the noise predictor ϵ_θ [9].

Attribute Encoder: The Attribute Encoder E_A is a down-sampling network that incorporates attention blocks and residual blocks. This type of down-sampling network is commonly used in both detection [22] and generation [23] tasks. It encodes the original image into a latent space through down-sampling. The latent embedding of visual features is denoted as $E_A(i_{\text{ref}}) \in \mathbb{R}^D$, which is obtained by taking the reference image as input.

Editing Generator: The Editing Generator generates the target image by inserting the fusion features Δh into the latent space of a *frozen* diffusion model. Previous studies have demonstrated the remarkable performance of diffusion models [23], [13] in this context. In this paper, we use DDIM [24] to train a *frozen* diffusion model as Editing Generator.

C. Step-Wise Alignment

In this section, we propose Step-Wise Alignment (SWA) to align the fusion prompt from the Attribute Encoder and Text Encoder. SWA optimizes the Attribute Encoder to align the latent embedding of the editing image with the reference image, guided by the embedding of the textual prompt.

A straightforward approach to perform latent manipulation during the generation of x_0 from x_T is to update the Attribute Encoder to minimize the following loss:

$$\mathcal{L}_{\text{dir}}(i_{\text{out}}, t_{\text{target}}; i_{\text{edit}}, t_{\text{source}}) := 1 - \frac{\Delta I \cdot \Delta T}{\|\Delta I\|\|\Delta T\|}, \tag{2}$$

Algorithm 1 Image Editing via SWA

```
Input: An editing image $i_{\text{edit}}$; A text prompt $t_{\text{attr}}$; Editing Generator; $\epsilon_\theta$;Visual Generator $G_V$; Attribute Encoder $E_A$; CLIP encoder $\xi_{clip}$; Diffusion model timestep $T$; Timestep $t_{\text{swa}}$
Output: A target image $i_{\text{out}}$
Initialize $t_{\text{source}}$ and $t_{\text{target}}$ based on $t_{\text{attr}}$.
Generate the reference image $i_{\text{ref}} = G_V(t_{\text{target}})$.
Encode $\Delta h = E_A(i_{\text{ref}})$ .
Get the noise image $x_0$ from $i_{\text{edit}}$ based on $\epsilon_\theta$.
for $i = 1, 2, \ldots, N$ do
    for $t = T, T-1 \ldots, 0$ do
        if $t > t_{\text{swa}}$ then
            $x_{t-1} = \sqrt{\alpha_{t-1}} P_t(\widetilde{\epsilon}_\theta(x_t, t)) + D_t(\epsilon_\theta(x_t, t)) + \sigma_t \epsilon_t$.
        else
            $x_{t-1} = \sqrt{\alpha_{t-1}} P_t(\epsilon_\theta(x_t, t)) + D_t(\epsilon_\theta(x_t, t)) + \sigma_t \epsilon_t$.
    $i_{\text{out}} \leftarrow x_0$ .
    Update the parameters of Attribute Encoder $E_A$ as Equation (5).
return $i_{\text{out}}$
```

where $\Delta I = E_A(i_{\text{out}}) - E_A(i_{\text{edit}})$ and $\Delta T = E_T(t_{\text{target}}) - E_T(t_{\text{source}})$, for the generated image i_{out}, the editing image i_{edit}, the target prompt t_{target}, and the source prompt t_{source}.

However, this approach might cause image distortion or erroneous manipulations, as observed in prior works [13], [25]. An alternative approach entails adjusting the noise ϵ_t^θ predicted by the network during each sampling iteration. In brief, we can reformulate the diffusion process of DDIM as follows:

$$x_{t-1} = \sqrt{\alpha_{t-1}} P_t(\epsilon_\theta(x_t, t)) + D_t(\epsilon_\theta(x_t, t)) + \sigma_t \epsilon_t, \quad (3)$$

where $P_t(\epsilon_\theta(x_t, t)) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \sqrt{1 - \alpha_t} \epsilon_\theta(x_t, t) \right)$ as the predicted x_0, and $D_t(\epsilon_\theta(x_t, t)) = \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \epsilon_\theta(x_t, t)$ as the direction to x_t. Nonetheless, making direct modifications to the noise ϵ_θ in both P_t and D_t leads to mutual nullification, yielding an unchanged $p_\theta(x_{0:T})$. This phenomenon mirrors a form of destructive interference, as elucidated in Asyrp [26, Theorem 1], which will inadvertently nullify the effects of optimizing ϵ_θ.

Hence, in order to circumvent the interference delineated in Equation (3), we adopt an asymmetrical form in SWA:

$$x_{t-1} = \sqrt{\alpha_{t-1}} P_t(\widetilde{\epsilon}_\theta(x_t, t)) + D_t(\epsilon_\theta(x_t, t)) + \sigma_t \epsilon_t. \quad (4)$$

Here, $\widetilde{\epsilon}_\theta(x_t, t)$ represents an adjustment to $\epsilon_\theta(x_t, t)$ grounded on the visual features Δh. This is achieved by introducing Δh into the original feature maps h_t derived from x_t.

The optimization of ϵ_θ to $\widetilde{\epsilon}_\theta$ is achieved by using the text prompt t to guide the generation process. Due to the absence of ground truth for editing images in editing tasks, fully supervised training methods are not applicable. Given that CLIP has the capability for vision-language alignment and can effectively evaluate the editing results without ground-truth images, we use CLIP loss to fine-tune the Attribute Encoder.

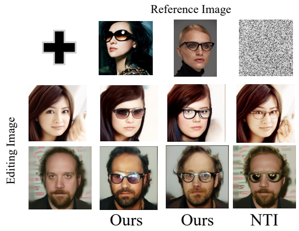


Fig. 2: **Consistent editing:** Once a reference image is provided, our method enables consistent and controllable editing to the reference image. In contrast, NTI fails to generate the corresponding style for the glasses.

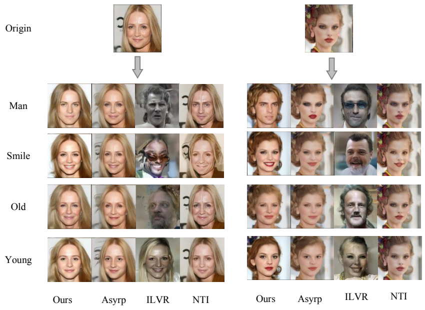


Fig. 3: **Editing results for in-domain attributes**

Following the approach presented in [27], we employ the directional CLIP loss in Equation (2) as our loss function:

$$\mathcal{L} = \lambda_{\text{clip}} \mathcal{L}_{\text{dir}}(\widetilde{P}_t, t_{\text{target}}; P_t, t_{\text{source}}) + \lambda_{\text{recon}} |x_{\text{out}}^t - x_{\text{edit}}^t|. \quad (5)$$

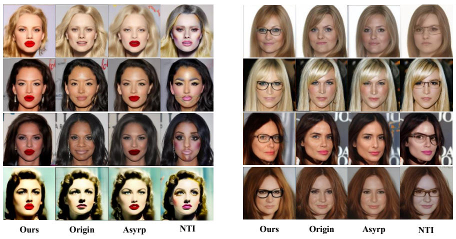
The modified $\widetilde{P}_t$ and the original P_t correspond to the formulations presented in Equation (4) and Equation (3), respectively. The last term in Equation (5) is the reconstruction loss, calculated as an ℓ_1 loss between the generated image and the original image. It effectively preserves the original features, preventing drastic alterations. To balance the aforementioned losses, we introduce the hyperparameters λ_{clip} and λ_{recon}.

D. Image Editing via SWA

Given an image $i_{\text{edit}} \in \mathbb{R}^{m \times n}$ and an attribute t_{attr}, the visual features from the reference image are integrated into the latent space of i_{edit} as detailed in Equation (4), and the textual prompt is used to optimize the Attribute Encoder as Equation (5). Meanwhile, the diffusion model (Editing Generator) remains *frozen*. The whole process is shown in Algorithm 1.

III. Experiments

Evaluation. Various metrics have been introduced in prior research to evaluate the effectiveness of image generation and editing. In this study, we employ the Inception Score (ISC) [28] and the Fréchet Inception Distance (FID) [29] as indicators of the image generation quality. Furthermore, we leverage the CLIP Score [20] to assess the alignment between edited images and their intended semantic targets.



(a) Makeup (b) Glasses

Fig. 4: **Editing results for out-of-domain attributes.**

TABLE I: In-domain Attributes Modification

		ILVR [13]	Asyrp [26]	NTI [25]	Ours
Man	ISC ↑	2.623	1.808	2.219	**3.292**
	FID ↓	150.2	77.65	52.23	**45.31**
	CLIP ↑	22.24	19.15	**23.02**	22.38
Old	ISC ↑	2.361	1.833	2.075	**2.778**
	FID ↓	145.1	77.82	**50.65**	55.96
	CLIP ↑	**23.26**	22.73	22.46	22.79
Smiling	ISC ↑	2.557	1.760	2.014	**2.573**
	FID ↓	223.7	73.93	**44.25**	85.36
	CLIP ↑	25.59	26.31	25.98	**27.01**
Young	ISC ↑	2.582	1.705	2.089	**2.624**
	FID ↓	147.9	80.03	**40.52**	55.28
	CLIP ↑	22.70	26.11	24.42	**25.07**

A. *Editing Consistency*

Our framework ensures high consistency between the attributes of editing images and their reference images. Although previous works, such as NTI [25], achieve realistic edits for real images, controlling the style of attributes remains challenging. As shown in Figure 2, SWA generates consistent styles of glasses across different images when given a specific reference image, whereas Null-Text Inversion cannot provide this level of control.

B. *Editing Generalization*

Both in-domain and out-of-domain attributes can be edited using our method. In-domain attributes refer to features that the *frozen* diffusion model has encountered during training. For instance, in the CelebA-HQ dataset, many images depict individuals with smiling expressions, and the attribute "smiling" is explicitly labelled in the dataset. On the other hand, out-of-domain attributes, such as "Add glasses", are not represented in the training data. As shown in Figure 3 and Table I, SWA demonstrates strong performance for in-domain attributes. Similarly, Figure 4 and Table II showcase SWA's ability to achieve high quality editing results for out-of-domain attributes as well. Figure 5 illustrates the visual results attained through SWA across various datasets.

C. *Ablation Experiments*

We depict the attribute editing process for the attribute "glasses" facilitated by SWA. Figure 6 effectively demonstrates that pixels are more *concentrated* on attribute features when a reference image is used (with SWA), highlighting the effectiveness of incorporating a reference image in generating visual features for the target attribute.

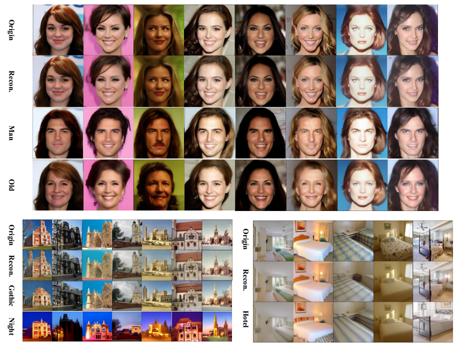


Fig. 5: **Editing results of SWA on various datasets.**

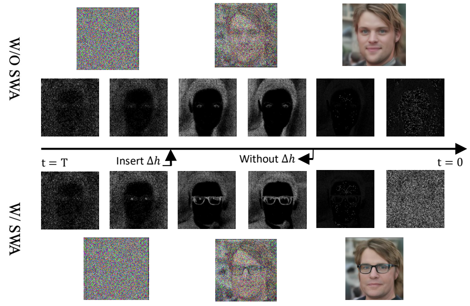


Fig. 6: **Ablation experiments of reference image:** The top half depicts the process without a reference image, while the bottom includes a reference image. Pixels are more *concentrated* on the attribute's features.

TABLE II: Out-of-domain Attributes Modification

		ILVR [13]	Asyrp [26]	NTI [25]	Ours
Makeup	ISC ↑	1.940	1.755	1.976	**2.467**
	FID ↓	144.4	77.49	**69.64**	88.0
	CLIP ↑	24.10	23.86	**26.46**	25.12
Glasses	ISC ↑	**3.060**	1.390	1.980	2.418
	FID ↓	165.6	147.5	**47.55**	124.6
	CLIP ↑	24.52	29.55	26.17	**30.07**

IV. Conclusion

This paper introduces SWA, a novel approach for manipulating real-world images by fusing textual and visual prompts. SWA integrates a blend of generated visual reference and textual guidance into the semantic latent space of a *frozen* diffusion model. By bridging the gap between visual patterns and textual semantics, SWA effectively alters both in-domain and out-of-domain attributes. In future work, our research will focus on improving the precision of attribute extraction from reference images. Specifically, we aim to refine methods for distinguishing attributes with similar visual features, thereby further improving the manipulation process.

References

[1] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *NeurIPS*, 2022, pp. 36 479–36 494.

[2] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *ICML*, 2021, pp. 8821–8831.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[4] OpenAI, "Gpt-4 technical report," 2023.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," pp. 1877–1901, 2020.

[6] Google, "Palm 2 technical report," 2023. [Online]. Available: https://ai.google/static/documents/palm2techreport.pdf

[7] B. Wallace, A. Gokul, and N. Naik, "Edict: Exact diffusion inversion via coupled transformations," *arXiv preprint arXiv:2211.12446*, 2022.

[8] G. Daras and A. G. Dimakis, "Multiresolution textual inversion," *arXiv preprint arXiv:2211.17115*, 2022.

[9] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," *arXiv preprint arXiv:2303.06555*, 2023.

[10] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[11] G. Kim, T. Kwon, and J. C. Ye, "Diffusionclip: Text-guided diffusion models for robust image manipulation," in *CVPR*, 2022, pp. 2426–2435.

[12] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut *et al.*, "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting," *arXiv preprint arXiv:2212.06909*, 2022.

[13] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," *arXiv preprint arXiv:2108.02938*, 2021.

[14] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *CVPR*, 2023.

[15] T. Nguyen, Y. Li, U. Ojha, and Y. J. Lee, "Visual instruction inversion: Image editing via visual prompting," *arXiv preprint arXiv:2307.14331*, 2023.

[16] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *ICLR*, 2021.

[17] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *CVPR*, 2022, pp. 11 461–11 471.

[18] X. Jia, Y. Zhao, K. C. Chan, Y. Li, H. Zhang, B. Gong, T. Hou, H. Wang, and Y.-C. Su, "Taming encoder for zero fine-tuning image customization with text-to-image diffusion models," *arXiv preprint arXiv:2304.02642*, 2023.

[19] Y. Nitzan, K. Aberman, Q. He, O. Liba, M. Yarom, Y. Gandelsman, I. Mosseri, Y. Pritch, and D. Cohen-Or, "Mystyle: A personalized generative prior," *TOG*, vol. 41, no. 6, pp. 1–10, 2022.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[21] T. Schick and H. Schütze, "Exploiting cloze questions for few shot text classification and natural language inference," *arXiv preprint arXiv:2001.07676*, 2020.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[23] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020, pp. 6840–6851.

[24] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[25] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," *arXiv preprint arXiv:2211.09794*, 2022.

[26] M. Kwon, J. Jeong, and Y. Uh, "Diffusion models already have a semantic latent space," *arXiv preprint arXiv:2210.10960*, 2022.

[27] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *CVPR*, 2022, pp. 18 208–18 218.

[28] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.