Appendix of "FreeDiff: Progressive Frequency Truncation for Image Editing with Diffusion Models"

Wei Wu[1,2], Qingnan Fan[2], Shuai Qin[2], Hong Gu[2], Ruoyu Zhao[3], and Antoni B. Chan[1]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong, China
weiwu56-c@my.cityu.edu.hk, abchan@cityu.edu.hk
[2] VIVO, Hangzhou, China
fqnchina@gmail.com, {shuai.qin,guhong}@vivo.com
[3] Xidian University, Xi'An, China royzhao@stu.xidian.edu.cn

A Preliminaries

Score-based diffusion models The diffusion process, characterized by multi-level noise perturbations, can be formulated as the discretization of Stochastic Differential Equation (SDEs) [20] and can be reversed if the scores of all noise levels are known. Different discretization formulations lead to different diffusion models [7, 18-20]. Denote the encoded image latent as $x_0 \in \mathbb{R}^{CHW}$. The objective of the denoising network ϵ_θ is to learn the score $\nabla_{x_t} \log p_{\sigma_t}(x_t|x_0)$ for the perturbed data x_t across all noise levels σ_t in the time step t [7, 19]:

$$\mathcal{L} = \mathbb{E}_t[w(t)\mathbb{E}_{x_0}\mathbb{E}_{x_t|x_0}[\|\epsilon_\theta(x_t) - \nabla_{x_t} \log p_{\sigma_t}(x_t|x_0)\|_2^2]] \tag{17}$$

where $w(t)$ is a positive weighting function, $\alpha_t \in (0, 1]$ is the noise schedule coefficient that controls the noise level and decreases to nearly 0 as t approaches T.

Guidance To influence the generation process via conditional distributions, we focus on $\nabla_{x_t} \log p_{\sigma_t}(x_t|c)$, where the condition c is the encoded embedding of the class labels, text prompt, etc. The conditional score [3] can be expressed as:

$$\nabla_{x_t} \log p_{\sigma_t}(x_t|c) = \nabla_{x_t} \log p_{\sigma_t}(x_t) + \nabla_{x_t} \log p_{\sigma_t}(c|x_t) \tag{18}$$

Classifier free guidance [8] is often used in T2I diffusion models as in Eq.2 and Eq.3, where $\epsilon_\theta(x_t, c)$ is the conditional score w.r.t. the encoded text prompt c, ϕ is the encoded embedding from a null (empty) string and $\epsilon_\theta(x_t, \phi)$ is its corresponding unconditional score. It is common practice to enlarge the guidance by a scaling factor $\gamma > 1$ since $p^\gamma(c|x_t) \propto p(x_t|c)/p(x_t)$, which equals to enhancing the posterior probability.

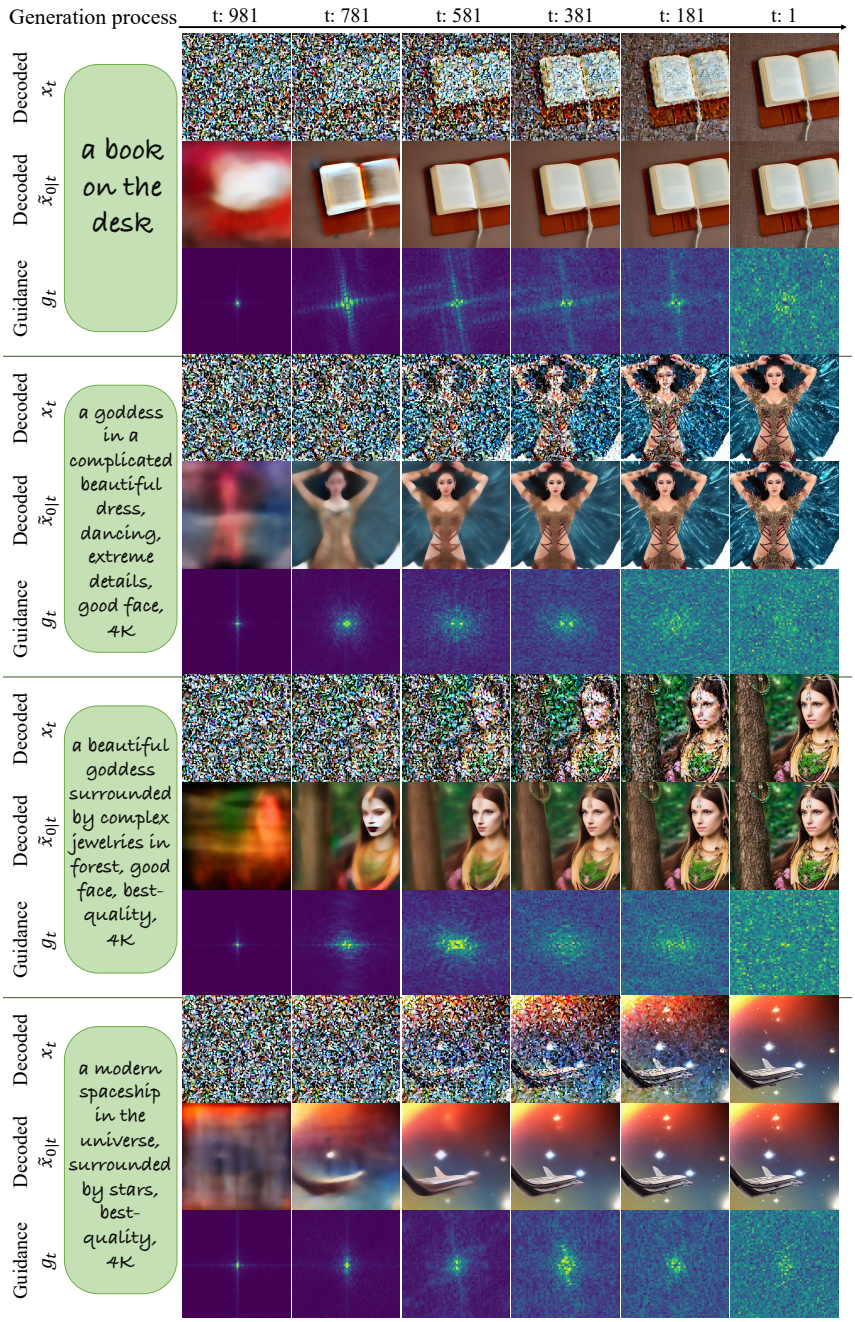

Fig. S1: Visualized decoded intermediate features and Fourier transformed features from a generation process with SD v1.5 [17].

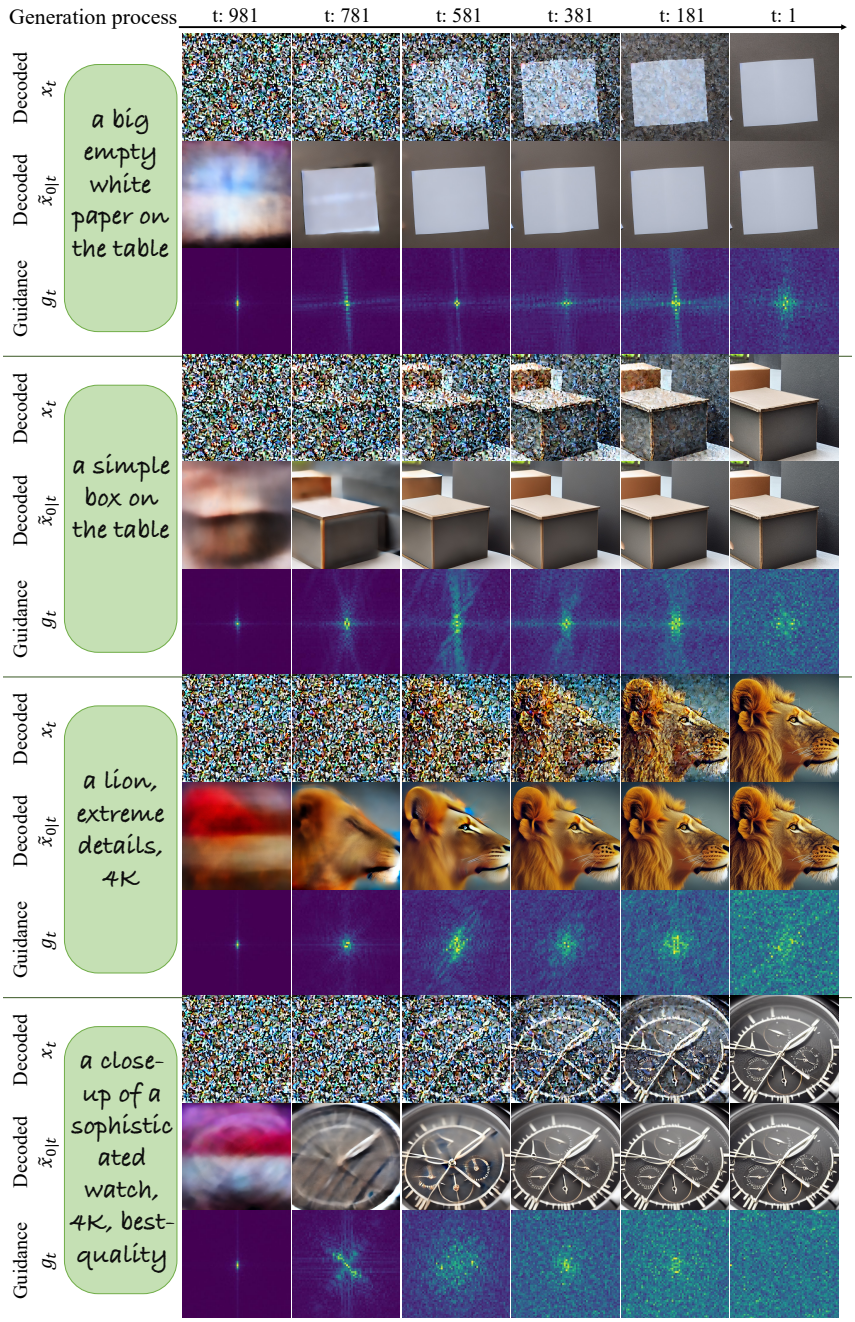

Fig. S2: Visualized decoded intermediate features and Fourier transformed features from a generation process with SD v1.5 [17].

B More intermediate features from generation process

More examples supporting our observations in the analysis section are listed in Fig. S1 and S2. The intermediate features are listed together with the prompt that generates the image. While these generated images show visual complexity of various levels, they follow a consistent generative pattern: details in $\tilde{x}_{0|t}$ are gradually added through the steps of generation, aligning with the gradual incorporation of higher frequency components from guidance.

C Editing difference from the frequency perspective

Examples of various editing types applied to different images using two ABMs, P2P [6] and PNP [1] , and direct editing are shown in Fig. S3. These examples support our hypothesis that direct editing inadvertently introduces an excess of low-frequency components, due to the learned prior and weighting schedule of the denoising network, leading to an undesired alteration in non-target regions.

D Qualitative and quantitative results

In this section, we first point out the existing problems within the PIE dataset [10] , and then detail our categorization of editing types from the frequency perspective and provide a default hyperparameters set for reference. Both quantitative results and qualitative examples are listed.

D.1 PIE Dataset

The PIE [10] dataset is the first large-scale dataset containing 700 images for quantitatively evaluating editing results across different editing types, with masks of target objects or regions provided for background-foreground assessment. However, there are two significant problems within the PIE dataset:

1. **Incorrect categorization** Within each editing type, some text-image pairs are misclassified. For example, in the "change object" category that aims at changing the identities of objects, the image "112000000008.jpg" is with prompt-pair "a painting of two women walking on the beach"-"a painting of two women walking on the grass", which would more aptly fit the "change background" category. Multiple misclassified editing pairs can be found in each category, hampering the dataset's credibility.
2. **Ill-defined category** In addition to the misclassification problem, the "change content" category, which primarily contains changes to objects, poses, materials, styles, encompasses only a minor portion of changes to shapes and expressions. These latter changes are more appropriate for the editing type "change content" to be distinguished from other types.

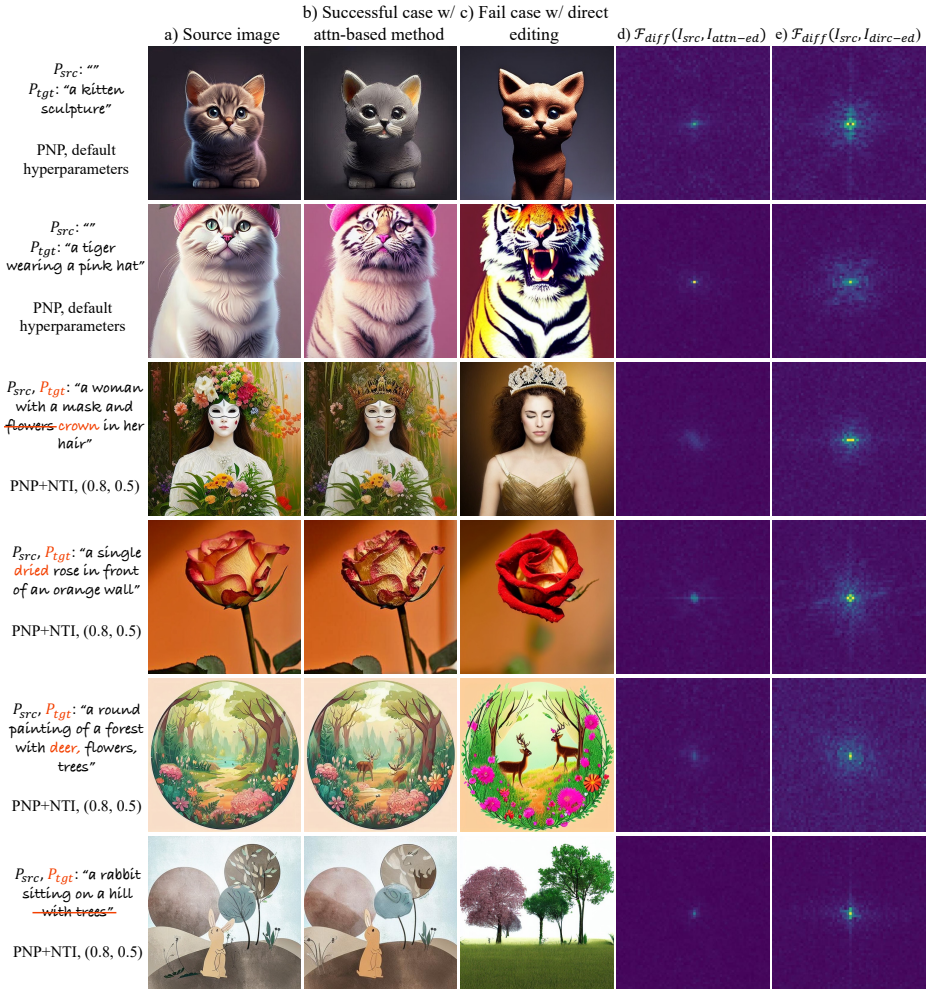

Fig. S3: Editing results from ABMs: P2P [6] +NTI [12] , PNP [1] +fixed-point inversion and directly applying guidance. Column d) and e) shows $\mathcal{F}_{diff}(I_{src}, I_{edit})$ between <source image, attention-based editing>, <source image, direct editing>, respectively. The $\mathcal{F}_{diff}(I_{src}, I_{edit})$ is normalized to the same numerical scale in each row.

These two problems hinder the accuracy of evaluation, since for ABMs, the best default hyperparameter sets vary largely for different editing types. For our method, accurately defining the editing types is crucial for selecting appropriate reference hyperparamters. Consequently, we will re-categorize the PIE dataset in the future, which will be detailed in the next section.

D.2 Editing categories and hyperparameters with *FreeDiff*

With *FreeDiff*, editing types are divided into three main categories from the frequency perspective:

1. *SF-0*: Changes that primarily rely on low-frequency components. This category includes editing types such as changing colors, environments, poses, shapes, and adding objects with significant structural differences compared to the original region. These changes require the alteration of low-frequency components and are affected during the earliest generation steps. In the situations of changing colors and environments, a two-step method is required to refine the guidance instead of directly applying frequency truncation.
2. *SF-1*: Changes that depend less on low-frequency components. Editing types in this category contain swapping object identities, removing an object, altering an object's material, changing style of the image, and adding objects. These changes rely less on low-frequency components and the editing mainly affects generation steps beyond the earliest.
3. *SF-2*: Changes that solely rely on high-frequency components. Editing types in this category are similar to the second type but focus on small objects as targets. These changes only rely on high-frequency components in later generation steps.

Given that our categorization primarily differentiates based on the spatial-frequency (SF) components involved, we denote these three categories as *SF-0*, *SF-1* and *SF-2*, respectively, for brevity consideration.

For notation simplicity, we consolidate T_{st}, T_{ed}, and τ_i by setting r_t^H to 32 outside the response period. With r_t^H set to 32 and given a guidance map of 64x64 dimensions, the high-pass filter will block all the signals and zero-out the guidance. For *SF-1*, one of the representative hyperparameter sets is $\{\tau_i = (781, 581, 1), r_t^H = (32, 10, 10)\}$, which means that we apply a high-pass filter with radii of 32, 10, and 10 for the time intervals $[981, 781]$, $(781, 581]$, and $(581, 1]$, respectively. As listed in Tab. S1, typical hypeparameter sets for *SF-1* are $\{\tau_i = (781, 581, 1), r_t^H = (32, 10, 10)\}$, $\{\tau_i = (781, 581, 1), r_t^H = (32, 32, 10)\}$, $\{\tau_i = (681, 581, 481, 1), r_t^H = (32, 20, 8, 1)\}$. For *SF-2*, typical hyperparameter sets are $\{\tau_i = (781, 581, 1), r_t^H = (32, 32, 20)\}$, $\{\tau_i = (781, 481, 1), r_t^H = (32, 32, 24)\}$. Notably, there are no typical hyperparameter sets for *SF-0*.

The choice of hyperparameter sets should primarily be based on the size of the object to be edited. We recommend using smaller high-pass filters in the earlier steps for editing larger objects.

D.3 *Two-step process* for editing colors and environments

To change colors and environments, we apply a *two-step process*. First, given that guidance truncated by *FreeDiff* at each step typically has smaller values on each pixel and a higher ratio of pixels that are activated within the target region, we aggregate the truncated guidance maps across all timesteps to form a coarse

Table S1: Hyperparameter sets for *SF-0*, *SF-1* and *SF-2*

SF Category	Hyperparameters	
	τ_i	r_t^H
SF-0	N/A	N/A
SF-1	(781, 581, 1)	(32, 10, 10)
	(781, 581, 1)	(32, 32, 10)
	(681, 581, 481, 1)	(32, 20, 8, 1)
SF-2	(781, 581, 1)	(32, 32, 20)
	(781, 481, 1)	(32, 32, 24)

mask for the target region. Then, we generate the target image by amplifying the guidance with this coarse mask, enhancing the refinement of the guidance. To preserve objects while changing the environment, the coarse mask can be reversed by subtracting it from a mask of ones. Some example results from this *Two-step process* are demonstrated in Fig. S5.

D.4 Quantitative results

For the overall quantitative results on the partial PIE dataset shown in Tab.1, we selected editing types where the comparison attention-based methods(ABMs) tended to perform well (change, add, and delete objects, change materials and poses). We did not include image where the inversion failed, or the ABMs catastrophically failed. Additionally, for most categories, we chose the former half of image-text pairs for the partial dataset. We did not cherry pick the images to improve our method's results. When testing the ABMs, we found some methods had a high number of failure cases on the claimed specialized type. Finally, we did not include editing types (style and color change) where ABMs required a large search to fine-tune the hyperparameters, since this is infeasible due to some method's computational complexity and lack of guidelines for searching, if we want to compare the results with our best hyperparameters. In total, 208 images were selected.

For the sub-categories results shown in Tab.2, we further correct the partial dataset from the mentioned issues, and 203 images are selected.

Overall, our experiments are conducted fairly since we select the images that the comparison ABMs perform well on. All editing results and hyperparams will be released with the source code.

We evaluated the CLIP score across the entire image and the LPIPS score for the background region, with results detailed in Tab. **??**. While our method exhibits slightly better over other ABMs, we do not consider the CLIP score to be an effective metric for evaluating editing quality. This is because, according to human perception, the editing results produced by P2P are generally better than those by PNP.

D.5 Qualitative results

A wide range of examples across all editing types in the figures attest to the effectiveness of our proposed method. For changes in materials, see Fig. S4; for changing styles, colors, and environments, see Fig. S5; for removing objects, see Fig. S6; for adding objects, see Fig. S7; for changing in identities, see Fig. S8.

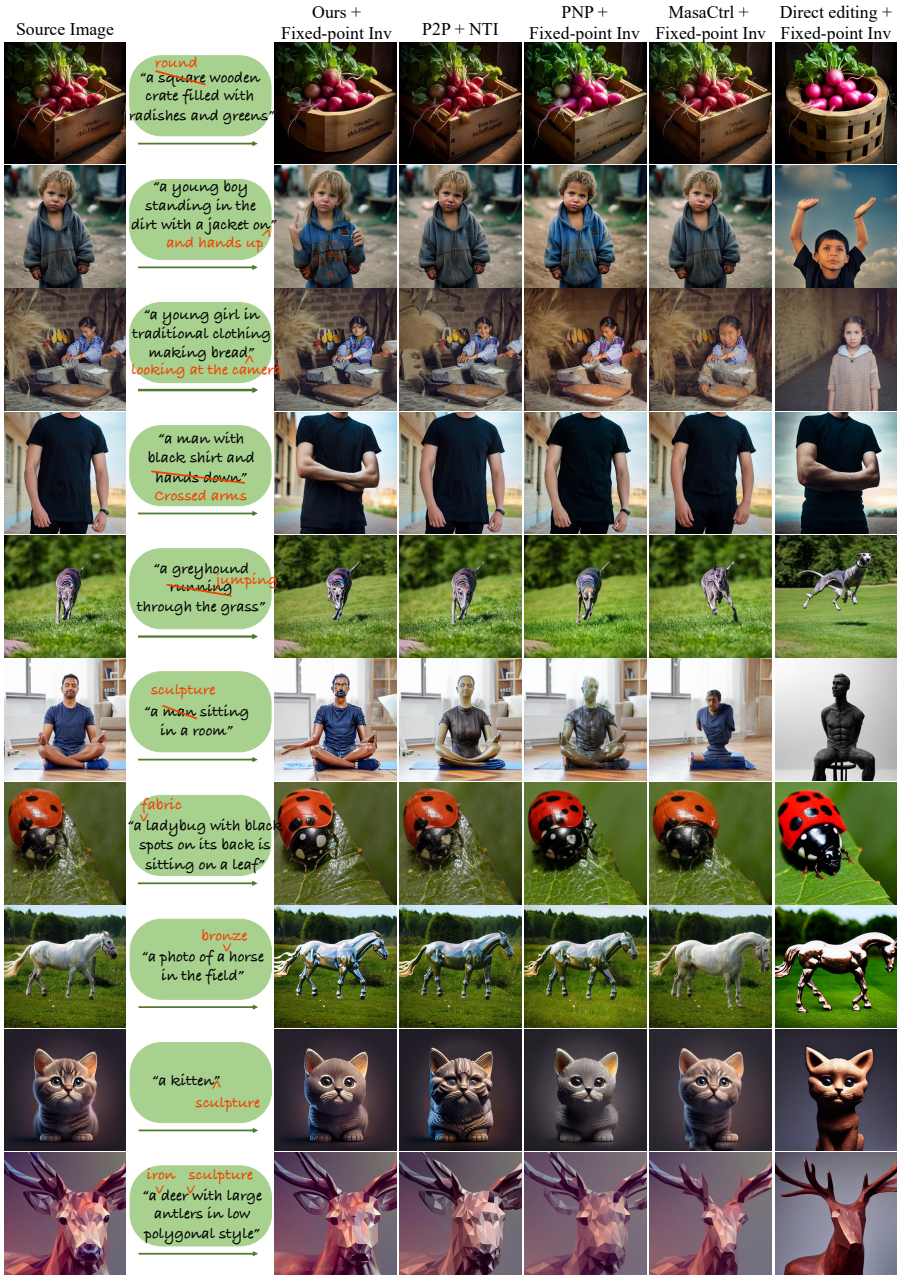

Fig. S4: Qualitative comparisons in changing materials, altering poses, and shapes using images from the PIE dataset [10] . The analysis juxtaposes our approach with 3 typical ABMs: P2P, PNP, and MasaCtrl. Direct editing results with fixed-point inversion are also included as a baseline for benchmarking.

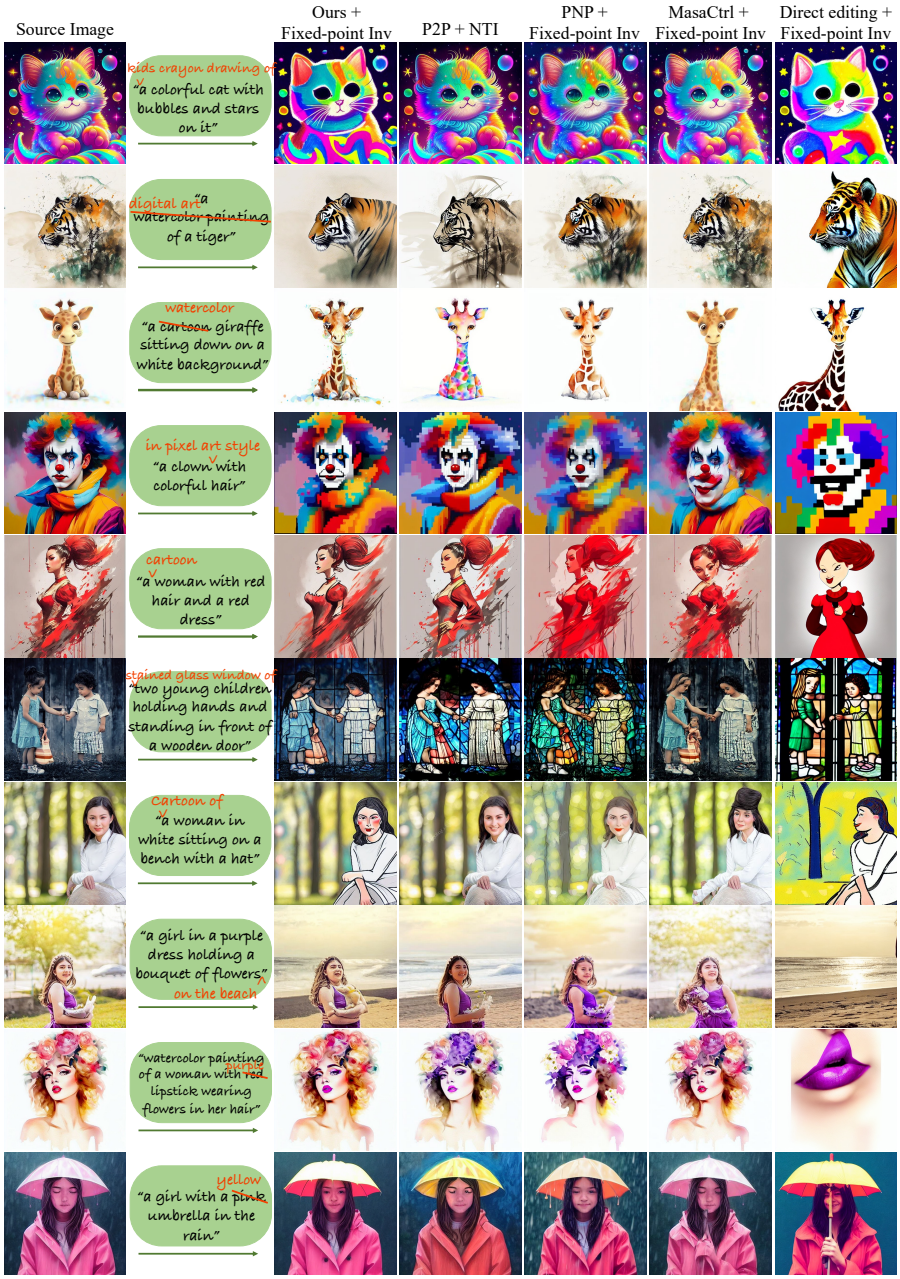

Fig. S5: Qualitative comparisons in changing styles, colors, and environment using images from the PIE dataset [10] . The analysis juxtaposes our approach with 3 typical ABMs: P2P, PNP, and MasaCtrl. Direct editing results with fixed-point inversion are also included as a baseline for benchmarking.

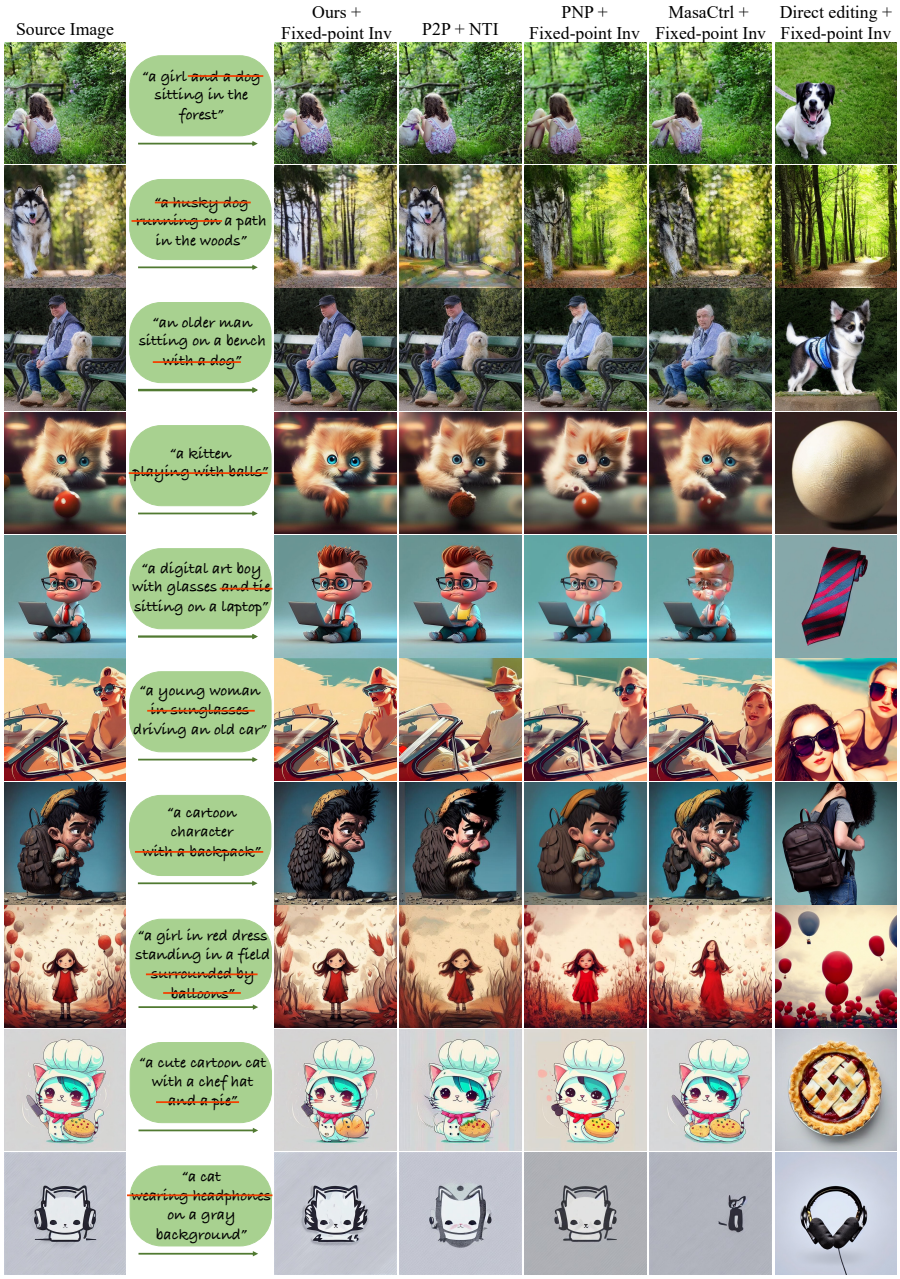

Fig. S6: Qualitative comparisons in removing objects using images from the PIE dataset [10] . The analysis juxtaposes our approach with 3 typical ABMs: P2P, PNP, and MasaCtrl. Direct editing results with fixed-point inversion are also included as a baseline for benchmarking.

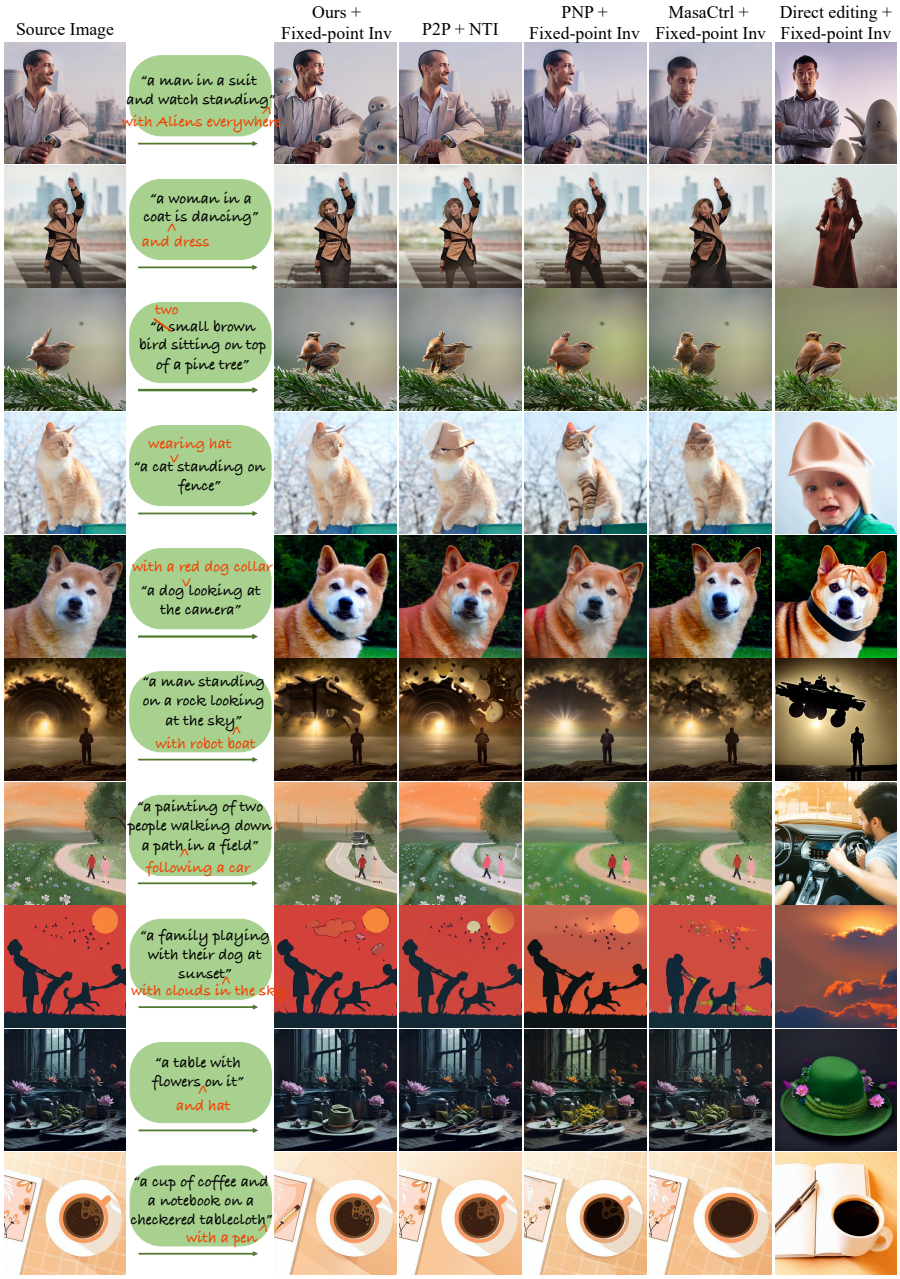

Fig. S7: Qualitative comparisons in adding objects using images from the PIE dataset [10] . The analysis juxtaposes our approach with 3 typical ABMs: P2P, PNP, and MasaCtrl. Direct editing results with fixed-point inversion are also included as a baseline for benchmarking.

Fig. S8: Qualitative comparisons in changing identities, altering shape, and adding objects using images from the PIE dataset [10] . The analysis juxtaposes our approach with 3 typical ABMs: P2P, PNP, and MasaCtrl. Direct editing results with fixed-point inversion are also included as a baseline for benchmarking.