Supplementary Materials for: NULL-text Inversion for Editing Real Images using Guided Diffusion Models

Ron Mokady* [1,2], Amir Hertz* [1,2], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or† [1,2]
[1]Google Research
[2]The Blavatnik School of Computer Science, Tel Aviv University

A. Evaluating Additional Editing Technique

Most of the presented results consist of applying our method with the editing technique of Prompt-to-Prompt [4]. However, we demonstrate that our method is not confined to a specific editing approach, by showing it improves the results of the SDEdit [7] editing technique.

In Fig. 1 (top), we measure the fidelity to the original image using LPIPS perceptual distance [13] (lower is better), and the fidelity to the target text using CLIP similarity [8] (higher is better) over 100 examples. We use different values of the SDEdit parameter t_0 (marked on the curve), i.e., we start the diffusion process from different $t = t_0 \cdot T$ using a correspondingly noised input image. This parameter controls the trade-off between fidelity to the input image (low t_0) and alignment to the text (high t_0). We compare the standard SDEdit to first applying our inversion and then performing SDEdit while replacing the null-text embedding with our optimized embeddings. As shown, our inversion significantly improves the fidelity to the input image.

This is visually demonstrated in Fig. 1 (bottom). Since the parameter t_0 controls a reconstruction-editability trade-off, we have used a different parameter for each method (SDEdit with and without our inversion) such that both achieve the same CLIP score. As can be seen, when using our method, the true identity of the baby is well preserved.

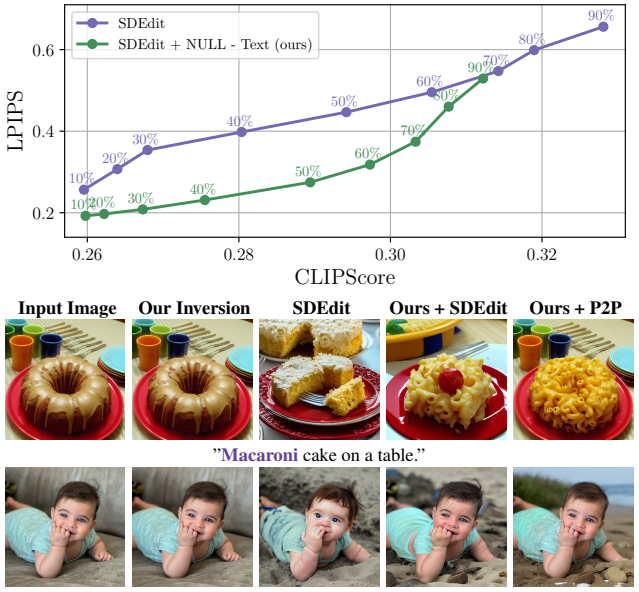


Figure 1. **Our method improves SDEdit results.** *Top: we evaluate SDEdit with and without applying NULL-text inversion. In each measure, a different SDEdit parameter is used, i.e., different percent of diffusion steps are applied over the noisy image (marked on the curve). We measure both fidelity to the original image (via LPIPS, low is better) and fidelity to the target text (via CLIP, high is better). Bottom, from left to right: input image, NULL-text inversion, SDEdit, applying SDEdit after NULL-text inversion, and applying Prompt-to-Prompt after NULL-text inversion. As can be seen, our inversion significantly improves the fidelity to the original image when applied before SDEdit.*

B. Limitations

While our method works well in most scenarios, it still faces some limitations. The most notable one is inference time. Our approach requires approximately one minute on GPU for inverting a single image. Then, infinite editing operations can be made, each takes only ten seconds. This is not enough for real-time applications. Other limitations come from using Stable Diffusion [9] and Prompt-to-Prompt editing [4]. First, the VQ auto-encoder produces artifacts in some cases, especially when human faces are involved. We consider the optimization of the VQ decoder as out of scope here, since this is specific to Stable Diffusion and we aim for a general framework. Second, we observe that the generated attentions maps of Stable Diffusion are less accurate compared to the attention maps of Imagen [10], i.e., words might not relate to the correct region, indicating inferior text-based editing capabilities. Lastly, complicated structure modifications are out of reach for Prompt-to-Prompt, such as changing a seating dog to a standing one as in [6]. Our inversion approach is orthogonal to the specific model and editing techniques, and we believe that these will be improved in the near future.

C. Societal Impact

Our work suggests a new editing technique for manipulating real images using state-of-the-art text-to-image diffusion models. This modification of real photos might be exploited by malicious parties to produce fake content in order to spread disinformation. This is a known problem, common to all image editing techniques. However, research in identifying and preventing malicious editing is already making significant progress. We believe our work would contribute to this line of work, since we provide an analysis of the inversion and editing procedures using text-to-image diffusion models.

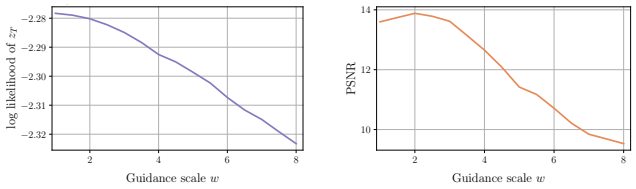
D. Ablation Study

Additional visual results for our ablation study are presented in Fig. 5 and 6, showing our method converges to high-quality reconstruction more efficiently. We now turn to provide additional results for specific experiments.

Robustness to different input captions. In Fig. 7 (top) we demonstrate our robustness to different input captions by successfully inverting an image using multiple captions. Yet, the edited parts should be included in the source caption in order to produce semantic attention maps for these (Fig. 7 bottom). For example, to edit the print on the shirt, the source caption should include a "shirt with a drawing" term or a similar one.

DDIM Inversion. To validate our selection of the guidance scale parameter of $w = 1$ during the DDIM Inversion (see Algorithm 1, line 3, in the main text), we conduct the DDIM inversion with different values of w from 1 to 8 using the same data as in Section 4. For each inversion, we measure the log-likelihood of the result latent image $z_T^* \in \mathcal{R}^{64\times64\times4}$ under the standard multivariate normal distribution. Intuitively, to achieve high edibility we would like to maximize this term since during training z_T^* distributes normally. The mean log-likelihood as a function of w is plotted in Fig. 2a. In addition, we measure the reconstruction with respect to the ground truth input image using the PSNR metric. As can be seen in Fig. 2b, increasing the value of w results in less editable latent vector z_T^* and poorer initial reconstruction for our optimization, and therefore we use $w = 1$.

Textual inversion with a pivot. We consider performing textual inversion around a pivot, i.e., similar to our pivotal inversion but optimizing the conditioned embedding. This results in a comparable reconstruction to ours, as demonstrated in Fig. 8 (bottom), but with poor editability. We analyze the attention maps (Fig. 8, top), observing that these



(a) Edibility (b) Reconstruction

Figure 2. **Setting the guidance scale for DDIM.** *We evaluate the DDIM inversion with different values of the guidance scale. On left, we measure the log-likelihood of the latent vector z_T with respect to multivariate normal distribution. This estimates the editability as z_T should ideally distribute normally and deviation from this distribution reduces our ability to edit the image. On right, we measure the reconstruction quality using PSNR. As can be seen, using a small guidance scale, such as $w = 1$, results in better editability and reconstruction.*

Table 1. **Inference time comparison.** *We measure both inversion and editing time for different methods. SDEdit is faster than ours, as an inversion is not employed by default, but fails to preserve the unedited parts. Our method is more efficient than the rest of the baselines, as it provides accurate reconstruction with faster inversion time, while also allowing multiple editing operations after a single inversion.*

Method	Inversion	Editing	Multiple edits
VQGAN + CLIP	—	$\sim$ 1m	No
Text2Live	—	$\sim$ 9m	No
SDEdit	—	10s	**Yes**
Imagic	$\sim$ 5m	10s	No
Ours	$\sim$ 1m	10s	**Yes**

are less accurate than ours. For example, using our NULL-text optimization, the attention referring to "goats" is much more local, and attention referring to "desert" is more accurate. Consequently, editing the "desert" results in artifacts over the goats (Fig. 8, bottom).

NULL-text optimization without pivotal inversion. Optimizing the null-text embedding fails without the efficient pivotal inversion. This is demonstrated in Fig. 5 and 6, where the non-pivotal NULL-text optimization produces low-quality reconstruction (2nd row).

E. Additional results

Additional editing results of our method are provided in Fig. 3 and additional comparisons are provided in Fig. 9.

Inference time comparison. As can be seen in Tab. 1, SDEdit is the fastest since an inversion is not employed, but as a result, it fails to preserve the details of the original image. Our method is more efficient than Text2Live [1], VQ-

GAN+CLIP [3] and Imagic [6], as it provides an accurate reconstruction in $\sim$ 1 minute, while also allowing multiple editing operations after a single inversion.

Figure 3. **Additional real image editing results for our method.**

Comparison to Imagic Quantitative comparison to Imagic is presented in Fig. 4, using the unofficial Stable Diffusion implementation. According to these measures, our method achieves better preservation of the original details (lower LPIPS). This is also supported by the visual results in Fig. 11, as Imagic struggles to accurately retain the background. Furthermore, we observe that Imagic is quite sensitive to the interpolation parameter α, as a high value reduces the fidelity to the image and a low value reduces the fidelity to the text guidance, while a single value cannot be applied to all examples. Moreover, the authors of Imagic apply their method on the same three images, presented in Fig. 11, using the parameters $\alpha = 0.93, 0.86, 1.08$. This results in much better quality, however, still the background is not preserved, the model is sensitive to α, and fine-tuning per editing operation is required.

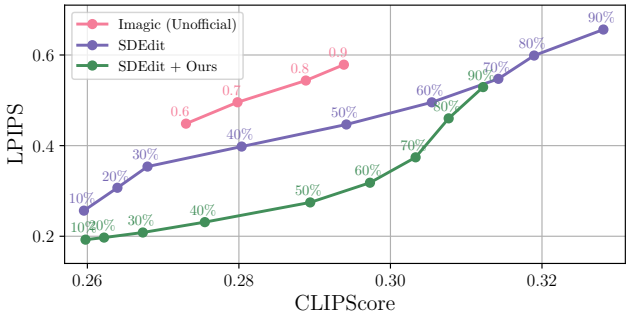


Figure 4. **Comparison to Imagic** *We quantitatively evaluate Imagic using the unofficial implementation for Stable Diffusion. We measure both fidelity to the original image (via LPIPS, low is better) and fidelity to the target text (via CLIP, high is better). Since the result is sensitive to the choice of the text embedding interpolation parameter α, we use different values, marked on the curve. The high LPIPS perceptual distance indicates that Imagic fails to retain high fidelity to the original image.*

F. Implementation details

In all of our experiments, we employ the Stable Diffusion [9] using a DDIM sampler with the default hyperparameters: number of diffusion steps $T = 50$ and guidance scale $w = 7.5$. Stable diffusion utilizes a pre-trained CLIP network as the language model ψ. The null-text is tokenized into *start-token*, *end-token*, and 75 non-text padding tokens. Notice that the padding tokens are also used in CLIP and the diffusion model since both models do not use masking.

All inversion results except the ones in the ablation study were obtained using $N = 10$ (See Algorithm 1 in the main paper) and a learning rate of 0.01. We have used an early stop parameter of $\epsilon = 1e-5$ such that the total inversion for an input image and caption took $40s - 120s$ on a single A100 GPU. Namely, for each timestamp t, we stop the optimization when the loss function value reaches $\epsilon = 1e-5$.

Baseline Implementations. For the comparisons in section 5, we use the official implementation of Text2Live* [1]

*https://github.com/omerbt/Text2LIVE

and VQGAN+CLIP[†] [2]. We have implemented the SDEdit [7] method over Stable Diffusion based on the official implementation[‡]. We also compare our method to Imagic [6] using an unofficial implementation[§] (see Appendix E).

Global Null-text Inversion. The algorithm for optimizing only a single Null-text embedding $\varnothing$ for all timestamps is presented in algorithm 2. In this case, since the optimization of $\varnothing$ in a single timestamp affects all other timestamps, we change the order of the iterations in Algorithm 1. That is, we perform N iterations in each we optimize $\varnothing$ for all the diffusion timestamps by iterating over t. As shown in Section 4, the convergence of this optimization is much slower than our final method. More specifically, we found that only after 7500 optimization steps (about 30 minutes) the global null-text inversion accurately reconstruct the input image.

Algorithm 2: Global NULL-text inversion

Input: A source prompt $\mathcal{P}$ and input image $\mathcal{I}$.
Output: Noise vector z_T and an optimized embedding $\varnothing$.

Set guidance scale $w = 1$;
Compute the intermediate results $z_T^*, \ldots, z_0^*$ of DDIM inversion for image $\mathcal{I}$;
Set guidance scale $w = 7.5$;
Initialize $\varnothing \leftarrow \psi(\text{""})$;
for $j = 0, \ldots, N-1$ **do**
  Set $\bar{z_T} \leftarrow z_T^*$;
  for $t = T, T-1, \ldots, 1$ **do**
    $\varnothing \leftarrow \varnothing - \eta \nabla_{\varnothing} \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \varnothing, \mathcal{C}) \right\|_2^2$;
    Set $\bar{z}_{t-1} \leftarrow z_{t-1}(\bar{z}_t, \varnothing, \mathcal{C})$;
  end
end
Return $\bar{z_T}, \varnothing$

G. Additional Background - Diffusion Models

Diffusion Denoising Probabilistic Models (DDPM) [5, 11] are generative latent variable models that aim to model a distribution $p_\theta(x_0)$ that approximates the data distribution $q(x_0)$ and easy to sample from. DDPMs model a "forward process" in the space of x_0 from data to noise. This is called "forward" due to its procedure progressing from x_0 to x_T. Note that this process is a Markov chain starting from x_0, where we gradually add noise to the data to generate the latent variables $x_1, \ldots, x_T \in X$. The sequence of latent variables, therefore, follows $q(x_1, \ldots, x_t \mid x_0) = \prod_{i=1}^{t} q(x_t \mid x_{t-1})$, where a step in the forward process is defined as a Gaussian transition $q(x_t \mid x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ parameterized by a schedule $\beta_0, \ldots, \beta_T \in (0, 1)$. When T is large enough, the last noise vector x_T nearly follows an isotropic Gaussian distribution.

An interesting property of the forward process is that one can express the latent variable x_t directly as the following linear combination of noise and x_0 without sampling intermediate latent vectors:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}w, \quad w \sim N(0, I), \tag{1}$$

where $\alpha_t := \prod_{i=1}^{t}(1-\beta_i)$.

To sample from the distribution $q(x_0)$, we define the dual "reverse process" $p(x_{t-1} \mid x_t)$ from isotropic Gaussian noise x_T to data by sampling the posteriors $q(x_{t-1} \mid x_t)$. Since the intractable reverse process $q(x_{t-1} \mid x_t)$ depends on the unknown data distribution $q(x_0)$, we approximate it with a parameterized Gaussian transition network $p_\theta(x_{t-1} \mid x_t) := N(x_{t-1} \mid \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$. The $\mu_\theta(x_t, t)$ can be replaced [5] by predicting the noise $\varepsilon_\theta(x_t, t)$ added to x_0 using equation 1.

Under this definition, we use Bayes' theorem to approximate

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\varepsilon_\theta(x_t, t)\right). \tag{2}$$

Once we have a trained $\varepsilon_\theta(x_t, t)$, we can using the following sample method

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t z, \quad z \sim N(0, I). \tag{3}$$

We can control σ_t of each sample stage, and in DDIMs [12] the sampling process can be made deterministic using $\sigma_t = 0$ in all the steps. The reverse process can finally be trained by solving the following optimization problem:

$$\min_\theta L(\theta) := \min_\theta E_{x_0 \sim q(x_0), w \sim N(0,I), t} \left\| w - \varepsilon_\theta(x_t, t) \right\|_2^2,$$

teaching the parameters θ to fit $q(x_0)$ by maximizing a variational lower bound.

H. User-Study

An illustration of our user study is provided in Fig. 12

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022. 2, 3

[2] Katherine Crowson. Vqgan + clip, 2021. https://colab.research.google.com/drive/1L8oL-vLJXVcRzCFbPwOoMkPKJ8-aYdPN. 4

[†] https://github.com/nerdyrodent/VQGAN-CLIP
[‡] https://github.com/ermongroup/SDEdit
[§] https://github.com/ShivamShrirao/diffusers/tree/main/examples/imagic

[3] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022. 3

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[6] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, abs/2210.09276, 2022. 1, 3, 4, 12

[7] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1, 4

[8] 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. *arXiv preprint arXiv:2103.00020*, 2021. 1

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 3

[10] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[11] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 4

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 4

[13] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1

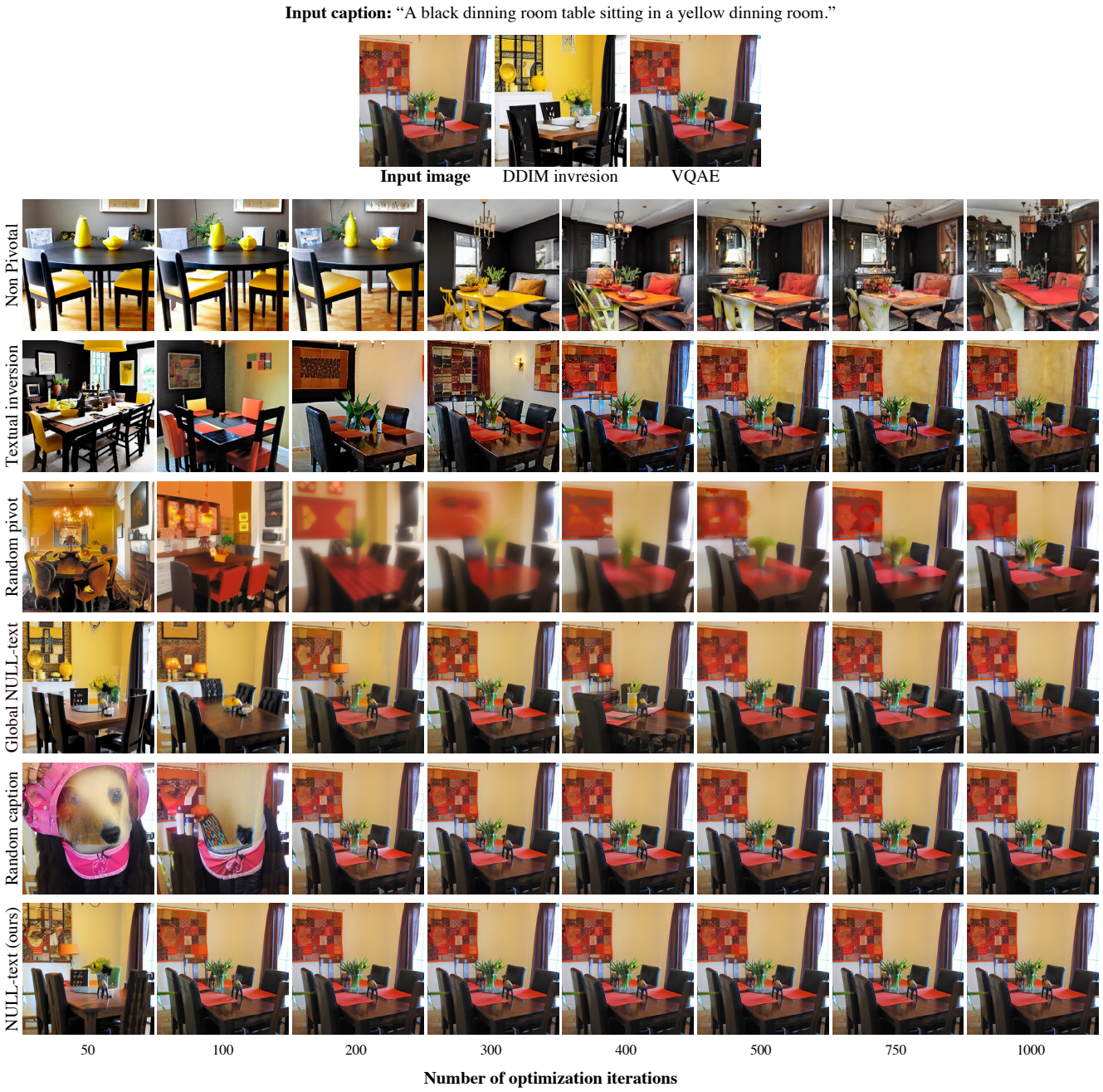


Figure 5. **Ablation study.** *We show the inversion results for an increasing number of optimization iterations. Our method achieves high-quality reconstruction with fewer optimization steps.*

Figure 6. **Ablation study.** *We show the inversion results for an increasing number of optimization iterations. Our method achieves high-quality reconstruction with fewer optimization steps.*

Figure 7. **Robustness to the input caption**. *We can invert an input image (top) using different input captions (first column). Naturally, the selection of the caption effects the editing abilities with Prompt-to-Prompt, as can be seen in the visualization of the cross-attention map (bottom). Yet, our method is not particularly sensitive to the exact wording of the prompt.*

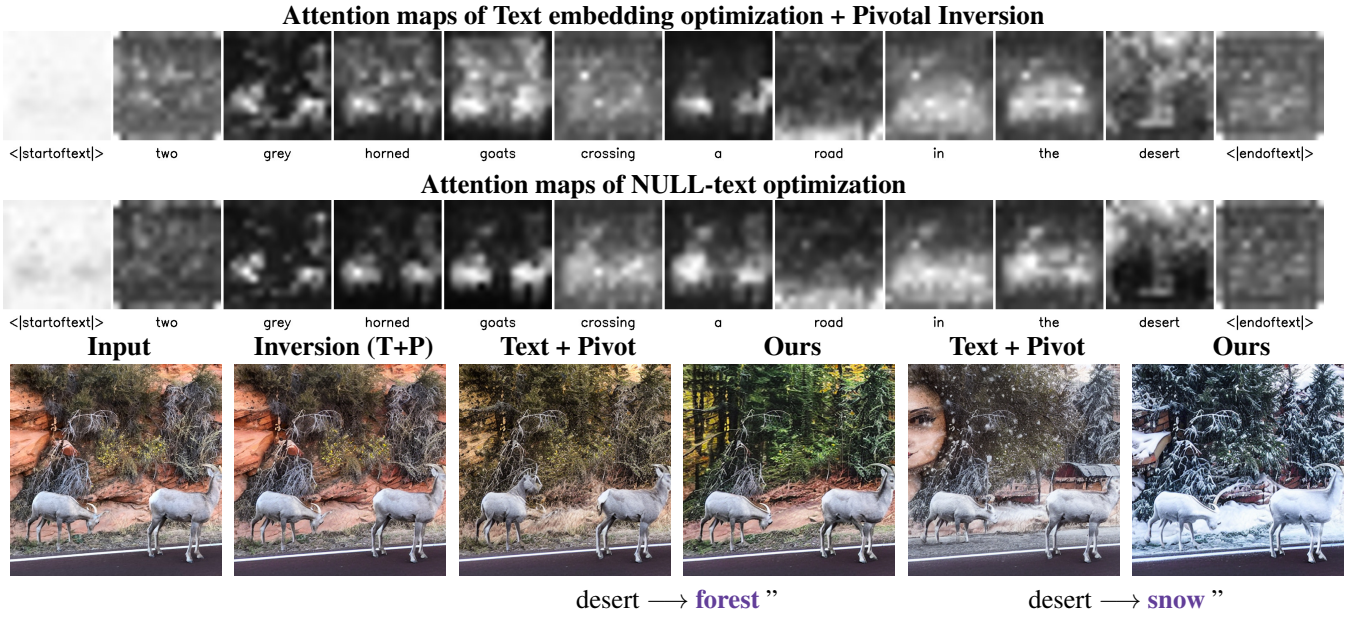


Figure 8. **Ablation study - Textual inversion with a pivot.** *We compare our method to replacing the text-NULL optimization with optimizing the conditional (textual) embedding while still applying pivotal inversion. As can be seen (top), this results in less accurate attention maps, and thus, in less accurate editing capabilities. In particular, textual inversion with a pivot achieves high-fidelity reconstruction (”Inversion (T+P)”), but goat heads distort (bottom) when editing is applied due to the inaccurate attention maps.*

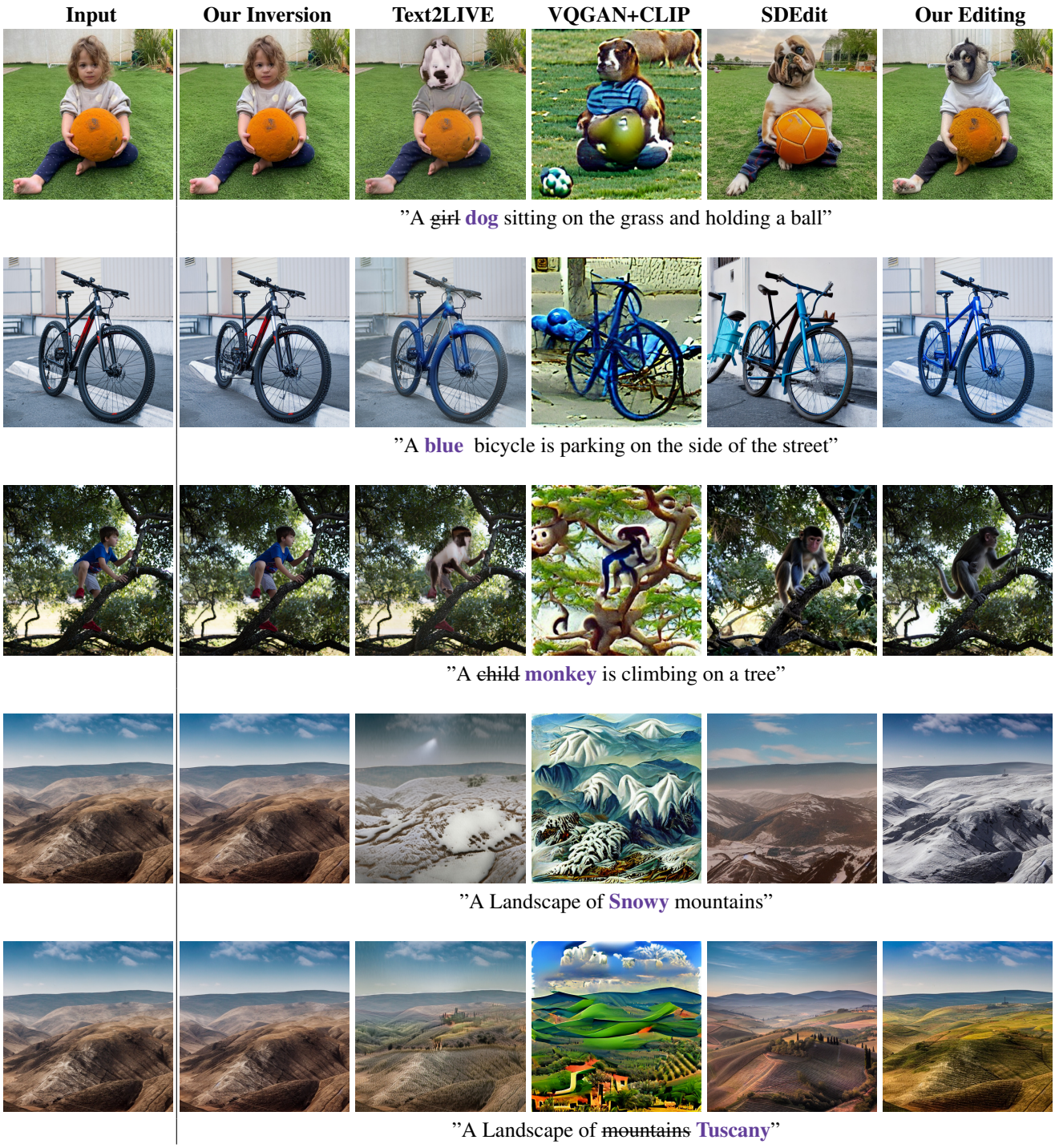


Figure 9. **Additional comparison results. See also Fig. 10**

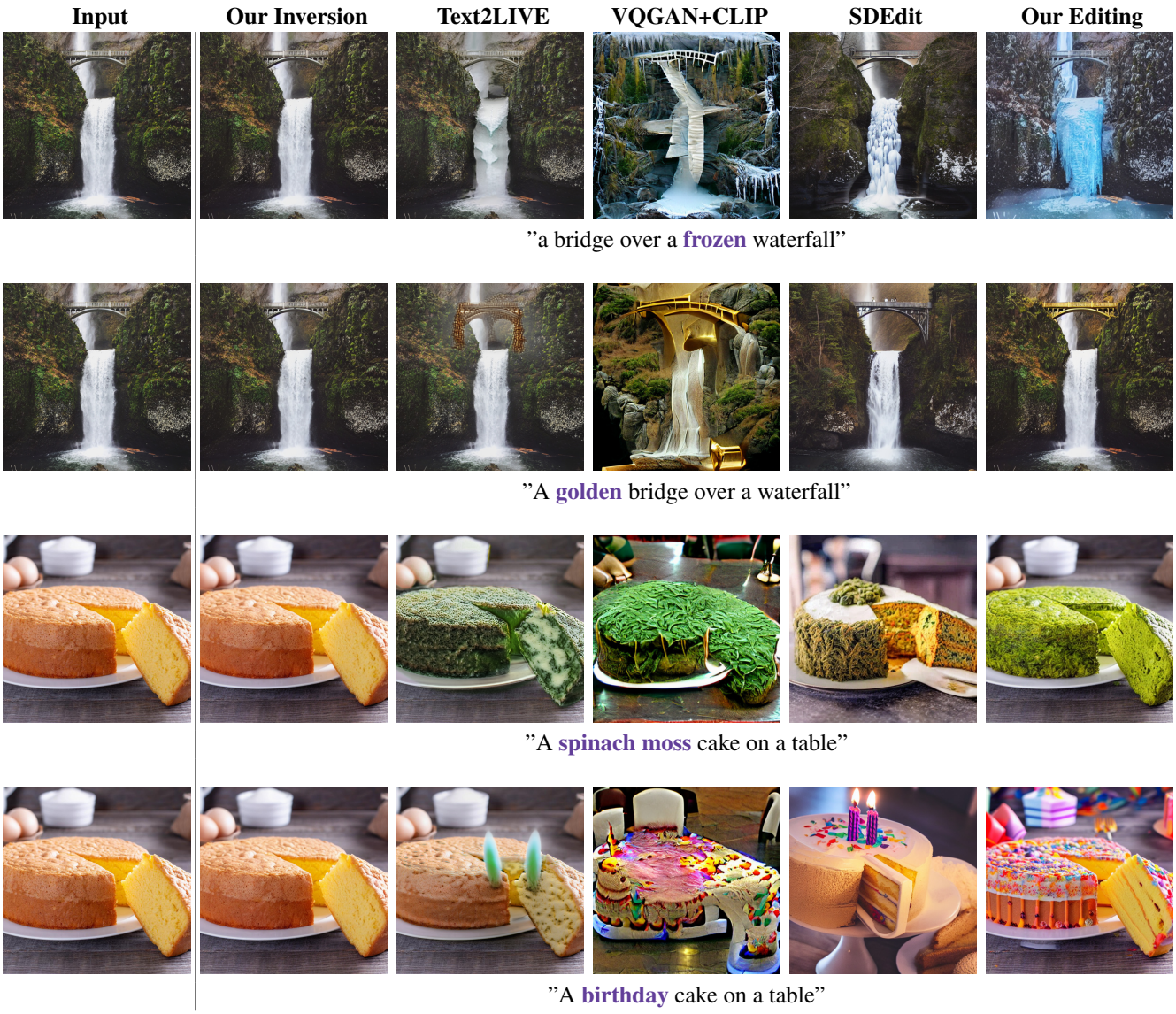


Figure 10. **Additional comparison results.**

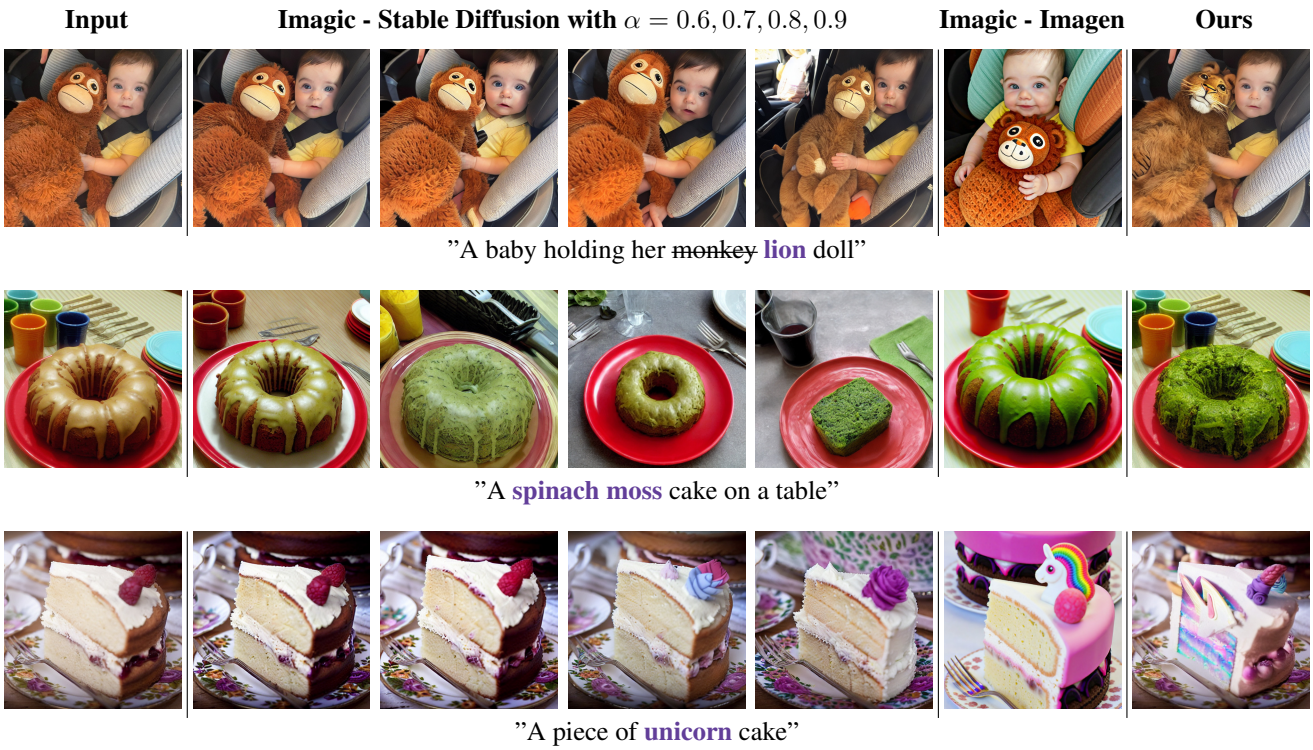


Figure 11. **Comparison to Imagic [6].** *We first employ the unofficial Imagic implementation for Stable Diffusion and present the results for different values of the interpolation parameter* $\alpha = 0.6, 0.7, 0.8, 0.9$ *(left to right). This parameter is used to interpolate between the target text embedding and the optimized one [6], where a larger value of* α *increases the fidelity to the target text. In addition, the Imagic authors apply their method using the Imagen model over the same images, using the following parameters* $\alpha = 0.93, 0.86, 1.08$ *(from top to bottom row). As can be seen, Imagic produces highly meaningful editing, especially when the Imagen model is involved. However, Imagic struggles to preserve the original details, such as the identity of the baby (*1*st row) or cups in the background (*2*nd row). Furthermore, we observe that each example requires a separate tuning of the* α *parameter. Lastly, recall that each Imagic editing requires a separate tuning of the model.*

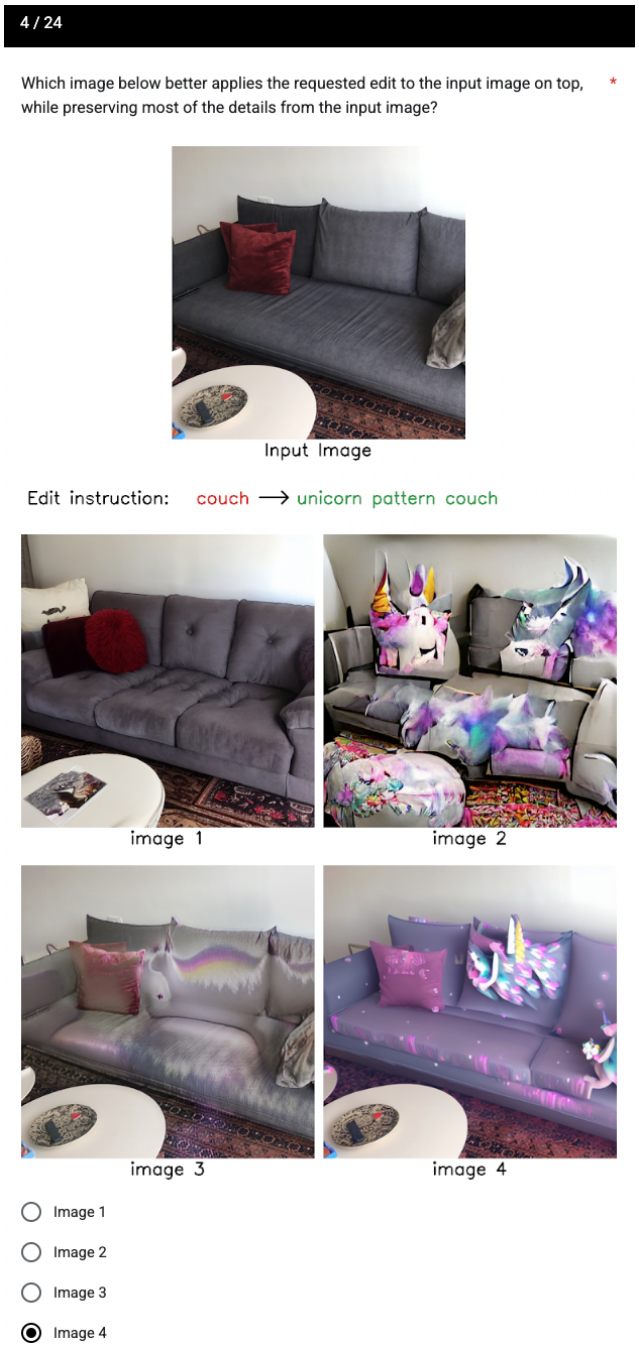


Figure 12. **User study print screen.**