

# NEAF: Natural Image Editing with Attention Fusion for Generalizable Test-time Optimization in Text-Guided Image Editing

## Supplementary Material

### 1. Feature injection

We incorporate feature injection across the processes within a triadic-feedback loop consisting of the source, edited, and reconstruction processes. Considering that the underlying diffusion model adopts a U-Net architecture [13], where the downsampling blocks primarily capture the coarse structure and layout of the latent representation, the editability with respect to the edit prompt  $p_e$  is governed by the text-to-image alignment capability of the diffusion model. Since the upsampling blocks are responsible for recovering fine-grained details, textures, and subtle visual attributes from the compressed latent, we adopt a feature injection strategy at the upsampling stages. Specifically, using null-text inversion [9], we obtain the initial noise vector  $z_T$  and the source caption prompt  $p_c$  from the source image. At each timestep  $t$  of the source reverse process, we inject the feature output from the source upsampling block, denoted as  $(f_t^c)$ , into the corresponding feature of the edited reverse process, denoted as  $(f_t^e)$ , as follows:

$$f_t^e \leftarrow f_t^e + f_t^c \quad (1)$$

Likewise, during the edited reverse process conditioned on the edit prompt  $p_e$  and noise vector  $z_T$ , we inject the edited upsampling feature  $(f_t^e)$  into the reconstruction reverse process’s upsampling block feature  $(f_t^r)$  at each timestep:

$$f_t^r \leftarrow f_t^r + f_t^e \quad (2)$$

Importantly, we do not apply feature injection throughout all diffusion timesteps. Since diffusion models such as Stable Diffusion [12] progressively refine the image from a coarse structure to fine details as the generation proceeds [15], only the latter 40 steps out of the total 50 timesteps are used for feature injection. The initial 10 timesteps are excluded from injection as these early steps primarily establish the overall structure and high-level semantic layout of the image based on the edit prompt  $p_e$ . Injecting source image features at this early stage could interfere with the model’s ability to adapt the coarse structure according to the desired semantic modifications. As the XA-Conductor alone is insufficient to fully preserve the fine visual details of the source image, feature injection plays a complementary role in enriching these details during the reverse processes. The XA-Conductor subsequently learns to accommodate and leverage these injected features within the feedback-driven editing framework.



Figure 1. **Results of ablation study on Feature injection.** Results showing that the removal of feature injection causes minor structural variations but does not significantly impact editability or source fidelity.

Fig. 1 presents the ablation results for feature injection. Even without feature injection as the default setting, there are cases where image editing performs well or even yields improved results, as shown in the dog image example in Fig. 1. As mentioned, feature injection helps preserve certain structural properties but does not significantly impact adherence to the edited prompt. This observation confirms that the XA-Conductor module alone is sufficient to identify editing-relevant cross-attention contribution levels, enabling effective image editing while maintaining source fidelity.

### 2. Ablation Studies

To address the contribution of each component, we conduct an incremental ablation study as shown in Fig. 2. While DDIM inversion [14] with BLIP-2 [8] reflects the target prompt, it fails to preserve the visual appearance of the source image. Replacing DDIM inversion with null-text inversion leads to better preservation of object appearance or background. Further incorporating BLIP-2 captioning improves text-image alignment, and the XA-Conductor with learned attention fusion consistently achieves the most faithful editing results, confirming the effectiveness of each component in our framework.

### 3. Additional qualitative results

To further validate our method against recent approaches, we conduct additional qualitative comparisons with LED-ITs++ (2024) [1], AURORA (2024) [6], FlowEdit (2025) [7] and SwiftEdit (2025) [10]. Following prior work, we evaluate on TEdBench [5] an image-text pair benchmark from Imagic which includes challenging non-rigid edit-

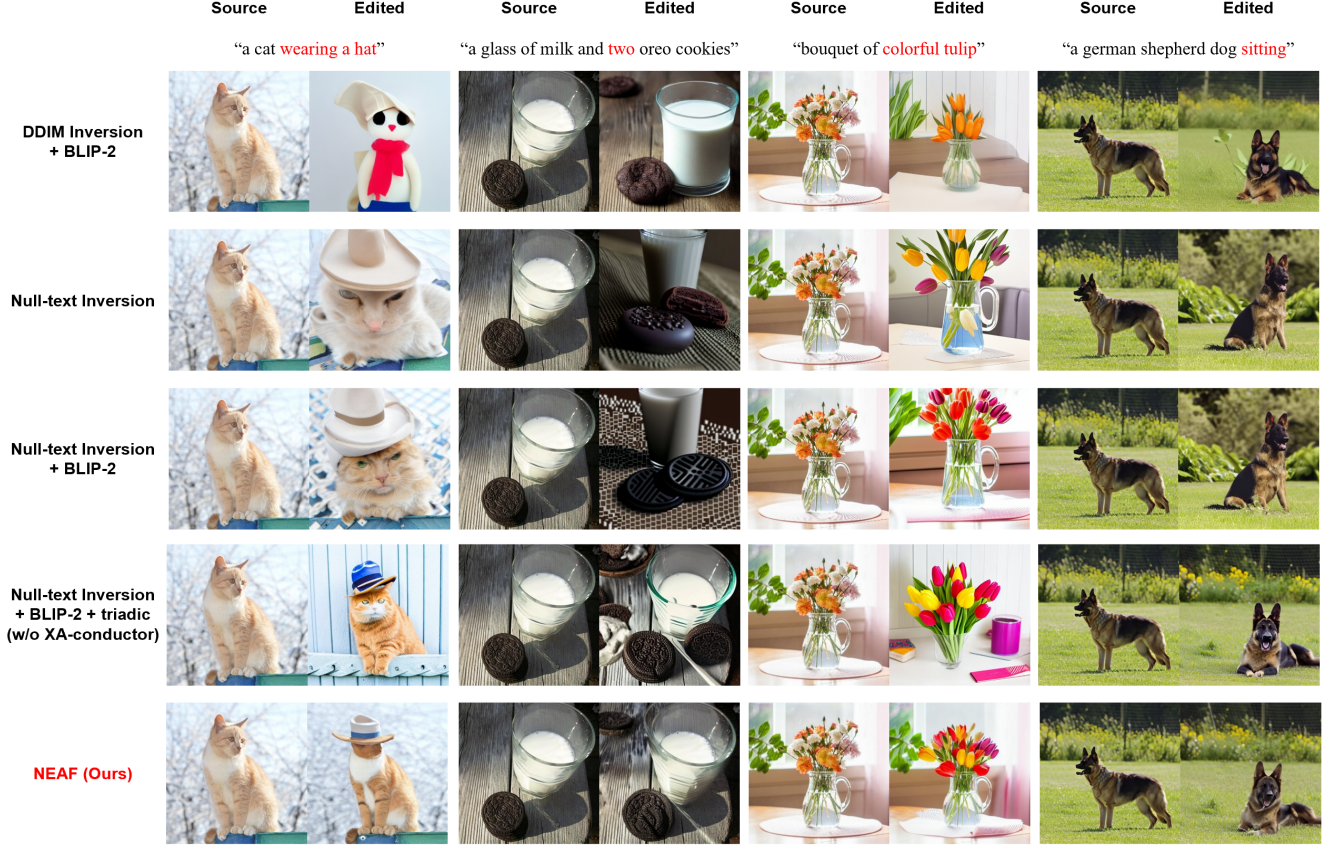


Figure 2. **Ablation study on individual components.** Each component progressively improves editing performance, demonstrating the contribution of each element to the final result.

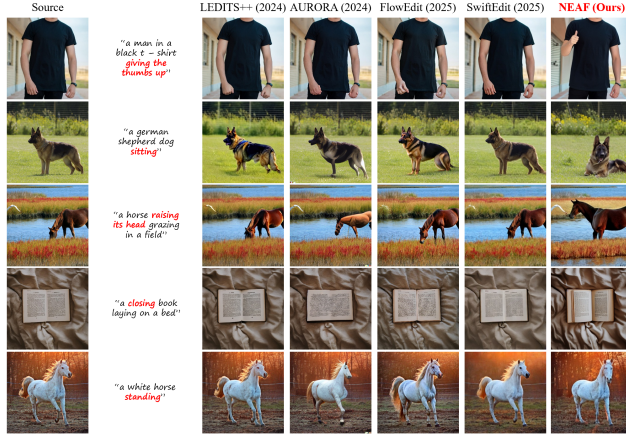


Figure 3. **Qualitative comparison with recent methods.** Our method demonstrates superior editability and source fidelity preservation compared to recent baselines.

ing scenarios such as complex semantic changes involving verb-driven transformations. As shown in Fig. 3, our method demonstrates accurate editing while effectively pre-

serving source image fidelity, particularly in non-rigid editing scenarios involving verb-driven semantic changes. Additional qualitative results of NEAF are presented in Fig. 4, showcasing its versatility across various editing prompts applied to the same source image.

## 4. User study details

NEAF applies non-rigid changes editing similar to Imagic, and as such no standard benchmark exists for evaluating non-rigid text-based image editing. Nevertheless, we conducted a quantitative evaluation using CLIP Score [4] and CLIP image similarity [2]. CLIP Score evaluates how well an image and text match based on a pretrained CLIP model [11]. CLIP image similarity, first utilized in InstructPix2Pix [2], measures how similar the edited image is to the source input image. However, we observed a trade-off between these metrics. Our NEAF ensures source fidelity while generating images that align with text prompts and achieves good average CLIP Scores. In contrast, models like InstructPix2Pix and PnP [15] preserve structure while changing style, score significantly higher on CLIP Similarity. This aligns with Imagic’s reasoning that LPIPS [16] and



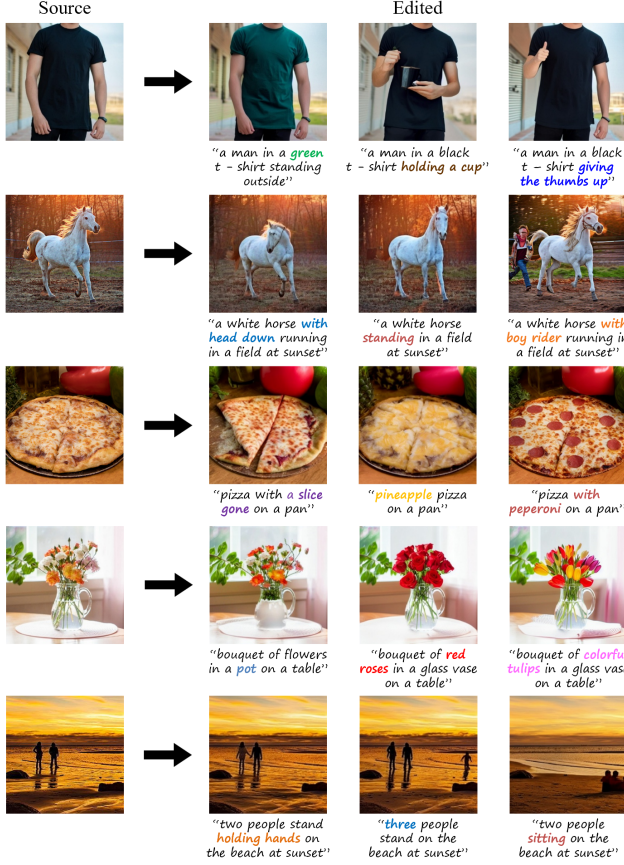


Figure 4. **Additional qualitative results of Different edited prompt applied to the same image.** NEAF generates different results for the same source image when given different edit prompt.

CLIP Score cannot faithfully assess an editing method’s performance [5].

To comprehensively evaluate performance on complex non-rigid image editing tasks beyond simple style transfer or color changes, we conducted a user study comparing 6 pairs of editing results with Imagic [5], PnP [15], and LED-ITs++ [1], respectively. Fig. 5 presents a screenshot of the user interface employed in our user study. To mitigate potential bias, we recruited participants from a broad demographic range spanning individuals in their 20s to 60s, with an approximate male-to-female ratio of 2.5:1 among the 50 participants. Each participant was presented with a total of 18 image pairs, consisting of 6 pairs comparing NEAF and Imagic, 6 pairs comparing NEAF and PnP, and 6 pairs comparing NEAF and LEDITs++. Each pair included a source image, an edit prompt, and the corresponding edited results. For fairness, we used the TEDBench dataset for Imagic, generated with hyperparameters carefully tuned to preserve as much source detail as possible while achieving the desired edits. The CFG scale was set to 5–7 for all baselines and NEAF. Notably, PnP and LEDITs++ preserve the surround-

## User study

**B** *I* U

(국문설명) 각 탭에는 원본 이미지 및 서로 다른 AI가 텍스트 프롬프트에 따라 수정된(edited) 2개의 이미지가 있습니다. 가장 잘 수정된 결과라고 생각하는 이미지를 선택해주세요. 평가를 위한 몇 가지 가이드라인은 다음과 같습니다:

- 더 사실적으로 보이는 수정 이미지 (부자연스럽거나 어색해 보이지 않음).
- 원본(소스) 이미지의 스타일과 디테일(배경이나 색상 등)을 유지하면서 텍스트 프롬프트에 충실한 수정 이미지

각 문항에 대해 2개의 수정된 이미지 중 하나를 선택해야 합니다. 총 12문항이 있습니다.

\*\* 구글 설문 특성상 그림이 작으므로, 수정 이미지를 상세히 들여다 보주시면 감사하겠습니다!

In each tab, we have a pair of images edited from the same original(source) image using different methods. Please select the radio button below the image you consider to be the best edited result. Here are some guidelines for evaluating good image editing:

- The edited image looks **more realistic** (not appearing unnatural or strange) .
- The edited image **matches with the TEXT PROMPT** better, while **preserving the style and details from the original(source) image**.

You must select one image from each pair of edited images for each question. There are 12 questions in total.

\*\* It would be great if you could check each editing carefully, as your answer matters a lot to us!

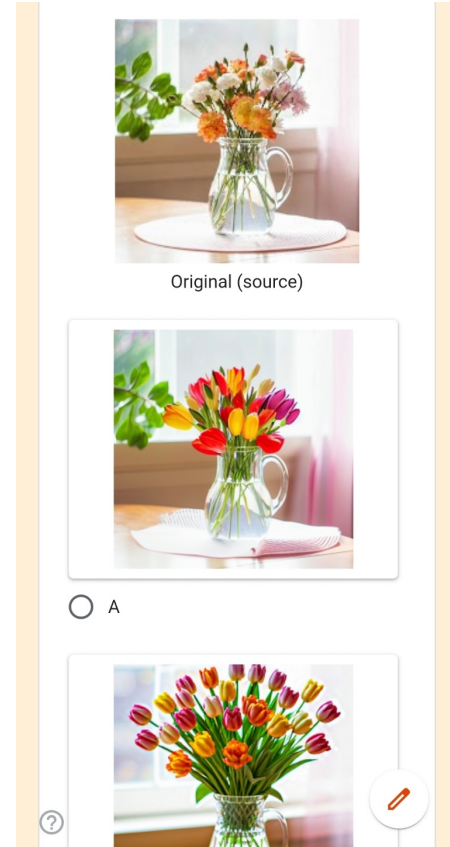


Figure 5. **User study screenshot.** A question format used in our human perceptual evaluation.

ing image content but fail to handle complex non-rigid edits, whereas Imagic is capable of performing such edits but at the cost of lower source fidelity. Our results demonstrate a strong user preference for NEAF, with over 70% of participants favoring NEAF’s outputs across all comparisons.

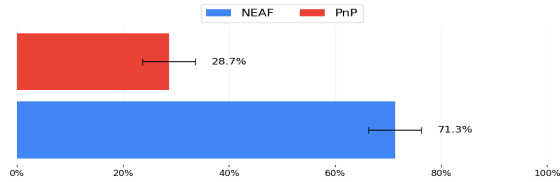


Figure 6. **User study result.** Preference rates for the image editing quality of NEAF in comparison to PnP.

Additional user preference results comparing NEAF with PnP are also presented in Fig. 6.

## 5. Limitations



Figure 7. **Failure cases.** Fidelity loss and partial edits in challenging scenarios.

We identify several failure cases of our method as shown in Fig. 7. In some cases, the edited result successfully follows the target prompt but fails to preserve the source image fidelity. In other cases, the edit is only partially applied despite accurate text alignment. As shown in the first row source fidelity is significantly degraded even when the edit is well-executed. These limitations partly stem from the constraints of the underlying Stable Diffusion v1.4 backbone and may be mitigated by adopting more recent generative models such as Stable Diffusion 3 [3]. In future work, we plan to extend our framework to handle more complex editing scenarios including multi-object interactions and fine-grained structural transformations.

## References

- [1] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *CVPR*, pages 8861–8870, 2024. 1, 3
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 2
- [3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yan-nik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 4
- [4] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2
- [5] Bahjat Kavar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023. 1, 3
- [6] Benno Krojer, Dheeraj Vattikonda, Luis Lara, Varun Jampani, Eva Portelance, Christopher Pal, and Siva Reddy. Learning Action and Reasoning-Centric Image Editing from Videos and Simulations. In *NeurIPS*, 2024. Spotlight Paper. 1
- [7] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. 1
- [8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. pages 19730–19742, 2023. 1
- [9] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 1
- [10] Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21492–21501, 2025. 1
- [11] 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. pages 8748–8763. PmLR, 2021. 2
- [12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021. 1
- [13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1
- [14] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1
- [15] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 1, 2, 3



- [16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. [2](#)