

Difference Inversion: Interpolate and Isolate the Difference with Token Consistency for Image Analogy Generation

Supplementary Material

1. Experimental Details

Implementation Details. When training the Difference tokens, we use a learning rate of 0.1 over 3000 steps, with a weight decay of 0.1. We also set the number of Difference tokens to 5 and the interpolation ratio α to 0.8, respectively. When obtaining the optimized prompts for images A and A' (prompt_A and prompt_B in Algorithm. 1, we follow the same settings as PEZ [10], using 10 tokens. For inference, we apply DDIM inversion [8] to generate a noisy x_t and then performed denoising to create B' . During this process, we set the start step to 45 out of 50 steps, though for cases requiring greater changes, up to 40 steps can be used. Additionally, we use a classifier-free guidance scale of 7.5.

Baselines. We compare Difference Inversion with three baselines: DIA [9], Analogist [3], and VISII [5]. Each baseline follows the original implementation as described in their respective papers. For DIA, we set the CLIP feature parameters s_c to 12.0 and the strength of the analogy s_t to 0.23684210526315788. In the case of Analogist, while the original paper utilizes the paid GPT-4V API, we instead used MiniCPM-V [11]¹, an open-source Vision Language Model (VLM) with similar performance to GPT-4V, due to its publicly accessible nature. For VISII, we use the same settings as those described in the original paper.

2. Additional Results

Ablation Test on λ_{tc} and λ_{clip} . We also conducted ablation experiments on the hyperparameters λ_{tc} and λ_{clip} in Eq. 10 for the Token Consistency Loss and CLIP Loss. Ultimately, we set the value of λ_{tc} to 0.01 and λ_{clip} to 6. The results for the scale of each hyperparameter can be found in Fig. 1 and Fig. 2, respectively.

More Complex tasks. We mainly use the InstructPix2Pix dataset to efficiently evaluate the $A : A' = B : B'$ relationship. Since the dataset contains many examples where A' images are generated from A images, we have adopted it as our main benchmark. However, our approach performs well not only when the correspondence between A and A' is explicit, but also when it is implicit, as in the case of DIA. Fig. 3 presents results for more complex examples in which the transformation from A to A' is implicit, meaning that A' is not directly generated from A .



Figure 1. **Visualization of the ablation study on λ_{tc} .** We finally use 0.01 as the value for λ_{tc} .

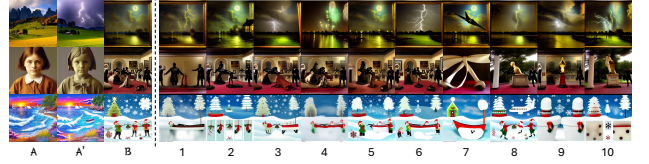


Figure 2. **Visualization of the ablation study on λ_{clip} .** We finally use 6 as the value for λ_{clip} .

Application of the same delta to multiple B . Since our approach extracts the difference between A and A' as discrete tokens, these tokens can be applied in a plug-and-play manner to multiple B images. We include additional experimental results demonstrating this capability in Fig. 4.

Qualitatively Results. Fig. 6 and Fig. 7 provide additional qualitative results. Furthermore, to demonstrate that our methodology is applicable not only to InstructPix2Pix dataset but also to a broader range of editing domains, particularly real-world datasets, we compare our approach with existing baselines using the *MagicBrush* [12] dataset. Note that, unlike InstructPix2Pix, the MagicBrush dataset contains only 1700 samples and has minimal overlap in text instructions, requiring us to manually select and pair B images to form triplets with A and A' . The results can be seen in Fig. 8 and Fig. 9.

3. Evaluation Details

Human Evaluation. For human evaluation, we survey a total of 60 participants with 50 examples. Each question is presented in a four-choice format, where participants are asked to select the most appropriate image as B' . The order of the options was randomized, and the format of the questionnaire can be found in Fig. 10.

VLMs Evaluation. We evaluate not only humans but also large-scale VLMs with strong reasoning capabilities. Un-

¹https://huggingface.co/openbmb/MiniCPM-Llama3-V-2_5

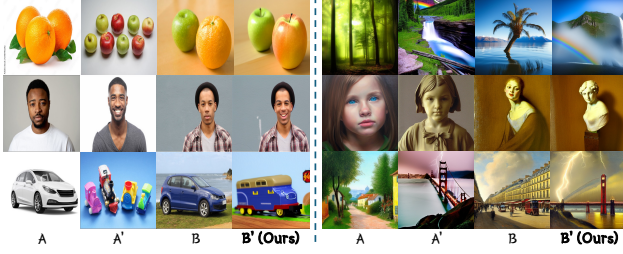


Figure 3. **More complex tasks.** *Left:* More complex examples used in DIA. *Right:* More complex examples generated using InstructPix2Pix.

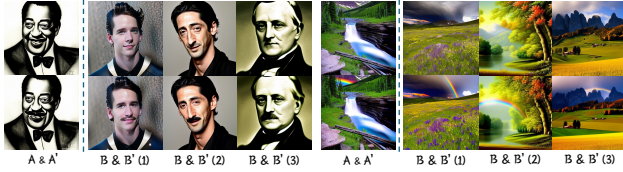


Figure 4. **Application of the same delta to multiple B .** Once a difference token is extracted from A and A' , it can be applied to multiple B images to generate their corresponding B' images.



Figure 5. **Colorizing the MNIST example in the form of an Image Analogy Formulation.** Despite coloring the digits 2 and 5 with identical RGB values, the CLIP similarity between $A \rightarrow A'$ and $B \rightarrow B'$ remains approximately 0.4, which is significantly lower than the ideal value of 1.

like human evaluation, we use a two-choice format, with the order of the options randomized. To assess whether large-scale VLMs are suitable as an evaluation metric, we verify if the model select the same answer when the order of the options for the same question is shuffled. Additionally, we examine the reasoning capability behind the chosen answer to ensure that the model understand the question correctly and select the right option. Detailed prompts and reasoning capabilities can be found in Fig. 11.

4. Inherited Limitations of CLIP space

Similar to many CLIP-based guidance methods [2, 4, 6], we utilize image and text embeddings in CLIP space [7] to extract the Difference between A and A' . However, as shown in the toy experiment in Fig. 5, we observe that even for an intuitive task like adding color to the MNIST [1] dataset, the cosine similarity between $A \rightarrow A'$ and $B \rightarrow B'$ is as low as approximately 0.4. Although we are able to refine the Delta through interpolation, the CLIP space demonstrated surpris-

ingly low similarity for accurate differences, contrary to human intuition. We anticipate that if a model with a better embedding space than CLIP emerges in the future, such a method could be explored further.

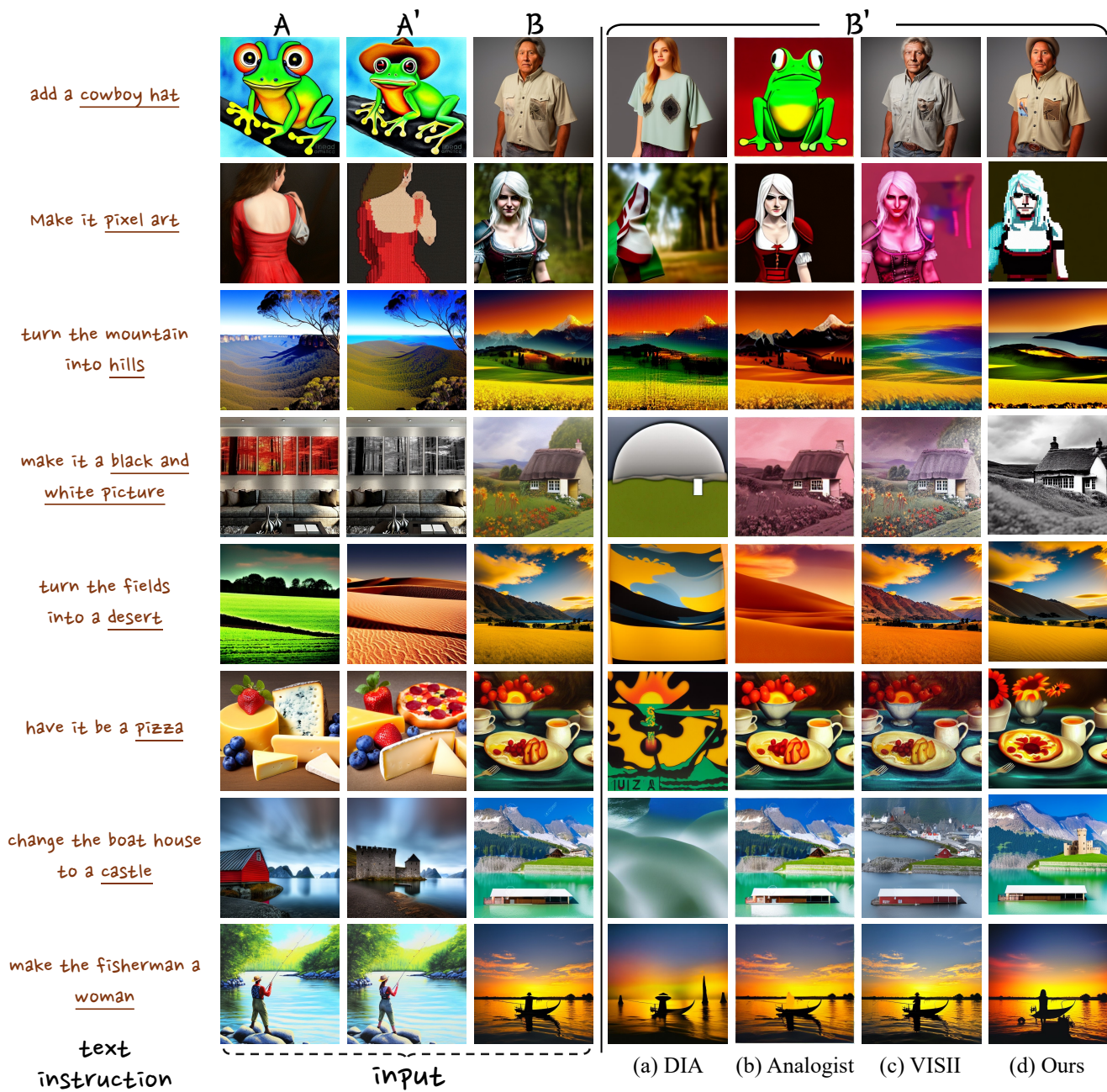


Figure 6. Additional qualitative comparison to baseline methods on the InstructPix2Pix dataset.



Figure 7. Additional qualitative comparison to baseline methods on the InstructPix2Pix dataset..

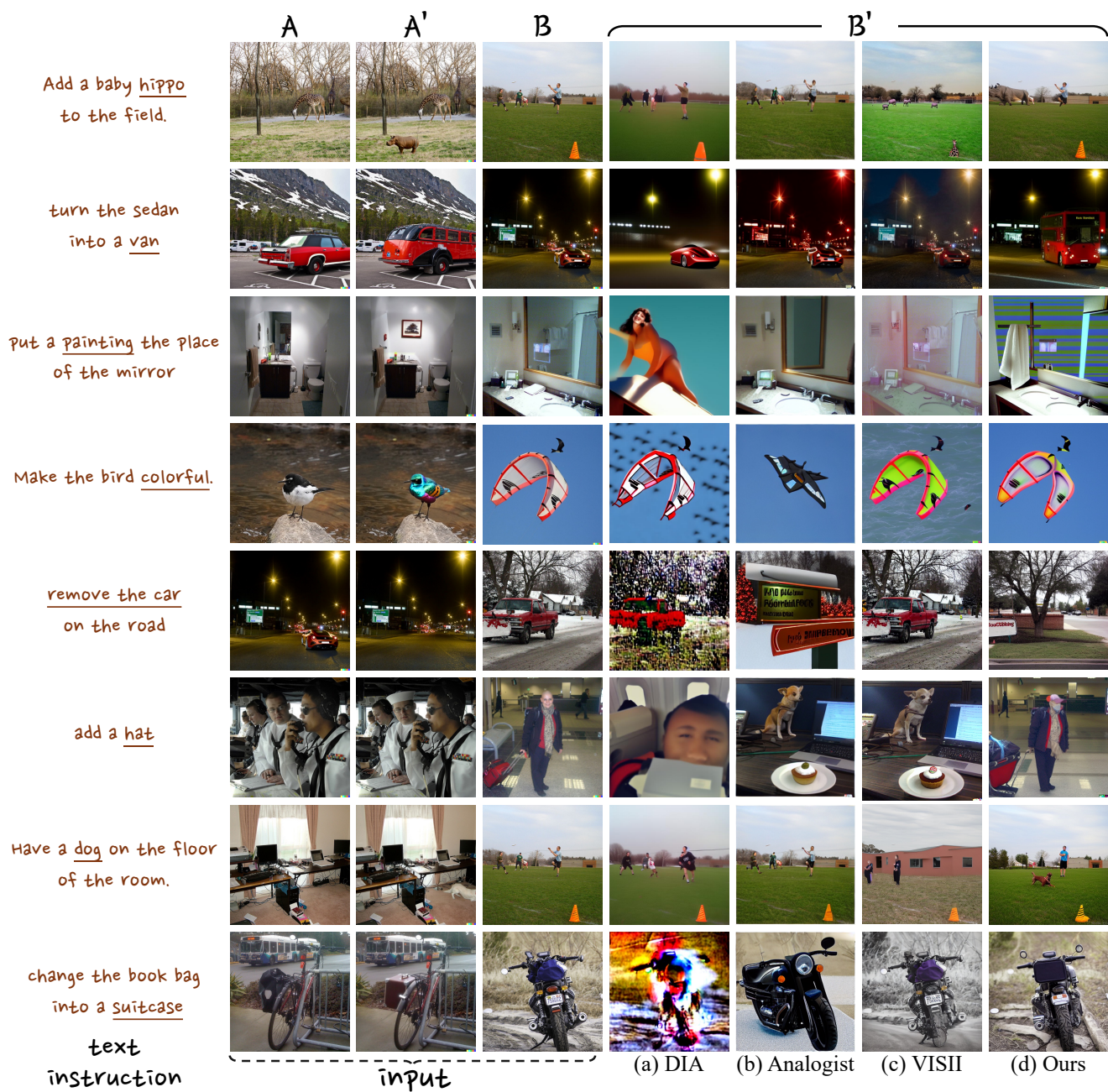


Figure 8. Qualitative comparison to baseline methods on the MagicBrush dataset.



Figure 9. Qualitative comparison to baseline methods on the MagicBrush dataset.

Visual Analogy Generation Human Evaluation

B *I* U  



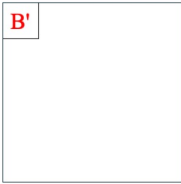
Given images A, A', and B, your task is to select the image that best represents B' from the following four options. You can answer with 1, 2, 3, or 4, based solely on which image most accurately reflects the transformation from B to match the relationship observed from A to A'. For example, in the example below, A and B are images of a tree and a tree struck by lightning, respectively. Since their difference is the presence of lightning, the correct B' would be an image of a house background with lightning added, corresponding to option 4. In this manner, please select the most suitable image for B' to complete the analogy A:A':B:???

(blank).


There are a total of 50 questions.

Example.


Visual analogy example:


A:  :: A':  :: B:  :: B': 

Options for B':

1: 

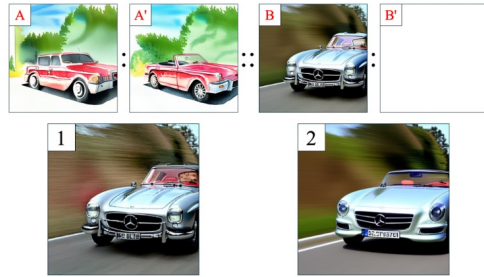
2: 

3: 

4: 

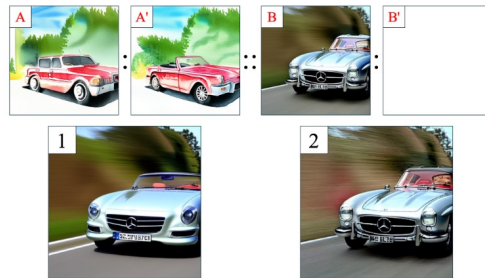
Option 4 is the correct answer, indicated by a checkmark.

Figure 10. Google Form Questionnaire for Human Evaluation.



Given images A, A', and B, your task is to select the image that best represents B' from the following two options. Choose either option 1 (left) or option 2 (right) based solely on which image most accurately reflects the transformation from B to match the relationship observed from A to A'. Answer only with '1' or '2'.

2



Given images A, A', and B, your task is to select the image that best represents B' from the following two options. Choose either option 1 (left) or option 2 (right) based solely on which image most accurately reflects the transformation from B to match the relationship observed from A to A'. Answer only with '1' or '2'.

1

Why?

The selection of option 1 is based on the observed transformation from A to A'—specifically, changing the vehicle from a closed-top model to a convertible (open-top) while maintaining other features like the overall color and style. Applying this transformation logic to B, option 1 represents B' with a convertible/open-top configuration while preserving the base design elements, thus reflecting the consistent transformation pattern observed in the relationship between A and A'. Option 2, while also depicting a similar car, does not follow the transformation pattern as closely, as it retains the closed-top appearance, contrary to the transition observed from A to A'.

Figure 11. Prompts and Reasoning Capability for VLM Evaluation.

References

- [1] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 2
- [2] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2
- [3] Zheng Gu, Shiyuan Yang, Jing Liao, Jing Huo, and Yang Gao. Analogist: Out-of-the-box visual in-context learning with image diffusion model. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 1
- [4] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 2
- [5] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [6] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 2
- [7] 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. 2
- [8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1
- [9] Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Šykora. Diffusion image analogies. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 1
- [10] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [11] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1
- [12] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 1