DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation Supplementary Material

Under-review

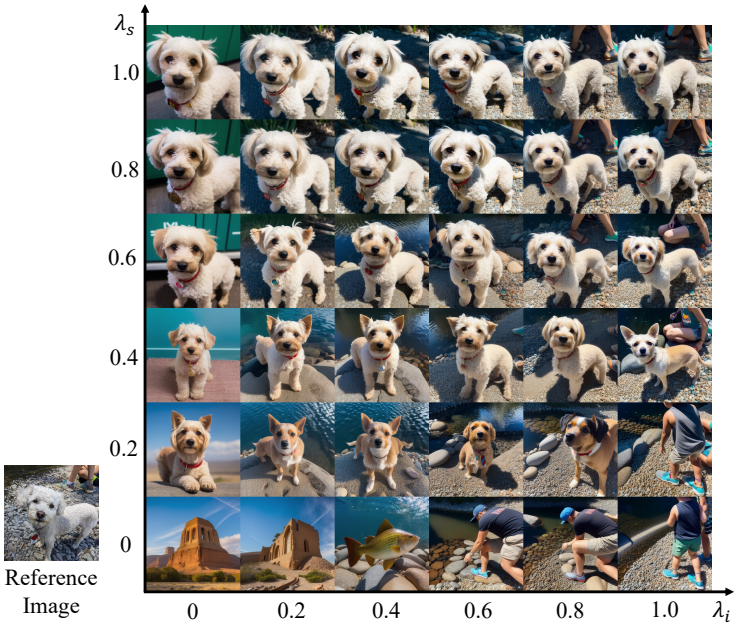
https://disenvisioner.github.io

A Effect of λ_s and λ_i

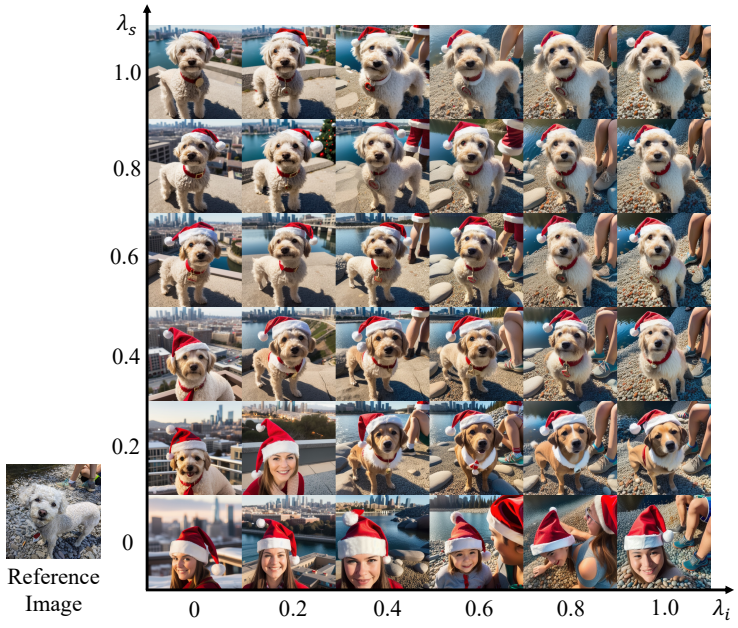
As defined in Eq. 5 of the main paper, the weights λ_s and λ_i serve to modulate the integration of information that is essential and irrelevant to the subject in the given visual prompt. To thoroughly assess their effect, we adjust their values systematically from 0 to 1.0 throughout the image customization process.

We generate images under varying settings of λ_s and λ_i employing both empty and non-empty (editing) prompts. Fig. 1 demonstrates that as λ_i decreases progressively (moving from the right to the left columns), the presence of subject-irrelevant disturbances in the images notably declines. Conversely, enhancing λ_s (moving from the bottom to the top rows) brings more pronounced consistency in subject identity between the reference and generated image. When both λ_s and λ_i are reduced to their minimum value, *i.e.*, $\lambda_s = 0$ and $\lambda_i = 0$, the generated images are solely influenced by the textual prompt, without incorporating any information from the reference image. To further explore the role of additional information in the prompt, we generate images with specific class names included in the prompts. As illustrated in Fig. 1b and Fig. 2, particularly in terms of identity consistency, no matter the category-guidance (for instance, specifying the class name "dog") is provided or not, it does not alter the customization quality. This indicates that our approach effectively deciphers and extracts subject-essential attributes from the reference image. By doing so, it renders additional semantic information redundant, which is then eliminated.

It is also evident that when image generation focuses exclusively on solely subject-essential features ($\lambda_s = 1.0$ and $\lambda_i = 0$) or solely on purely subject-irrelevant features ($\lambda_s = 0$ and $\lambda_i = 1.0$), the reproduction of the subject and the extraneous surrounding content is achieved independently, devoid of any interference from one another. This phenomenon confirms the proficiency of our DisEnvisioner in precisely segregating and enriching subject-essential and non-essential features. It highlights the DisEnvisioner's exceptional customization performance without the need for test-time tuning, and relying solely on a single reference image.



(a) With empty prompt.



(b) With non-empty prompt: "wearing a santa hat with a city in the background".

Fig. 1: Effect of varying λ_s and λ_i with empty and non-empty prompts without providing class names.

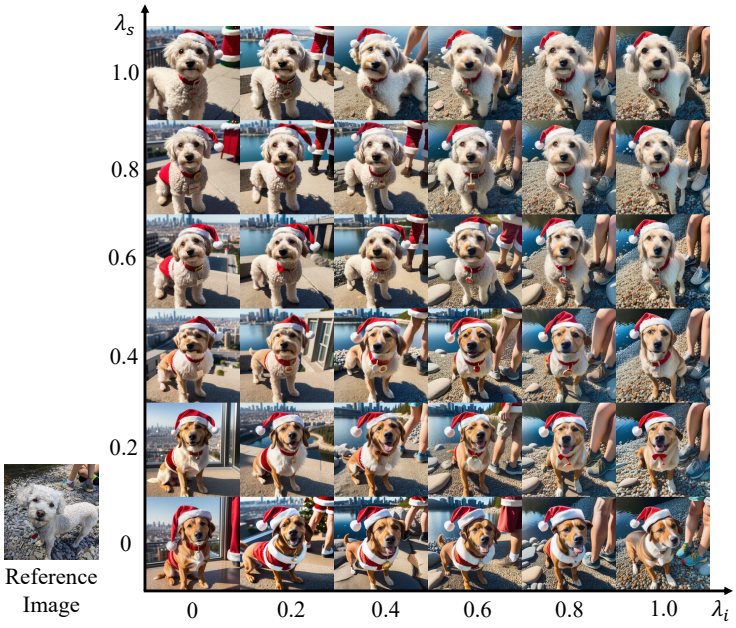


Fig. 2: Effect of varying λ_s and λ_i with providing class name in prompt. The prompt is "a **dog** is wearing a santa hat with a city in the background".

B Experimental Details

B.1 Grading of User Study

During each round of the user study, rather than **ranking** our DisEnvisioner and five other existing methods from the best to the worst, users are expected to **assign grades** from 0 to 5 to each method according to specific metrics. In practice, all participants have the complete freedom to grade any method with any score based on their personal judgment. After a total of 345 rounds of evaluation, the best-performing method often receives the highest scores, while scores for other methods are frequently identical due to similar customization quality. Additionally, it is uncommon for users to assign scores as low as 0 or 1. Although the grading differences among methods are not particularly large, DisEnvisioner consistently outperforms others competitors across all evaluation criteria.

B.2 Training Data

In our experiments, we utilize the *training set* of OpenImages V6 [1] as our training dataset. Based on this dataset, we construct {prompt, image} pairs for training. As depicted in Fig. 3, the training images are derived by cropping and resizing the raw images in accordance with the bounding box annotations. To

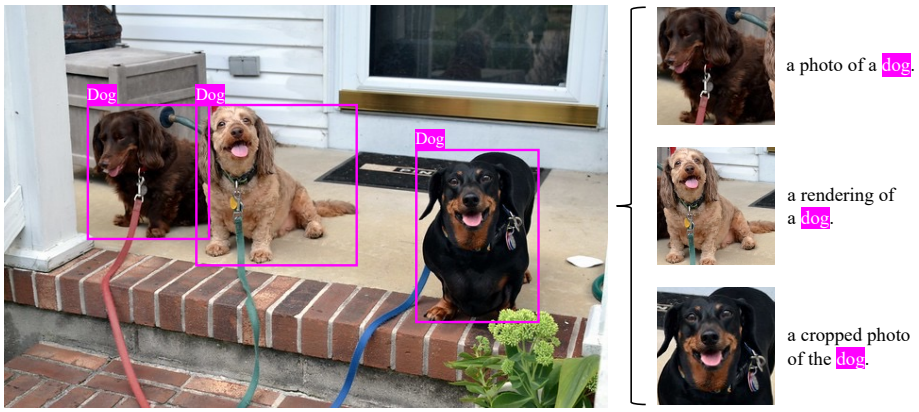


Fig. 3: Examples of training data. The training images are derived by cropping and resizing the raw images in accordance with the bounding box annotations.

ensure the quality of the training images, we further filter the cropped images: a cropped image is considered as unsatisfactory and therefore excluded if its area is greater than 80% or less than 2% of the original image's area. As a result, out of 14.61 million annotated bounding boxes, we obtain **6.82 million** {prompt, image} pairs. The text prompts are selected randomly from a CLIP ImageNet template [2] and integrated with labelled class names. The complete list of CLIP templates is provided below:

– "a photo of a S^*",
– "a rendering of a S^*",
– "a cropped photo of the S^*",
– "the photo of a S^*",
– "a photo of a clean S^*",
– "a photo of a dirty S^*",
– "a dark photo of the S^*",
– "a photo of my S^*",
– "a photo of the cool S^*",
– "a close-up photo of a S^*",
– "a bright photo of the S^*",
– "a cropped photo of a S^*",
– "a photo of the S^*",
– "a good photo of the S^*",
– "a photo of one S^*",
– "a close-up photo of the S^*",
– "a rendition of the S^*",
– "a photo of the clean S^*",
– "a rendition of a S^*",
– "a photo of a nice S^*",
– "a good photo of a S^*",

Fig. 4: Examples of testing data. We randomly selected 8 subjects of the 30, and three images are shown for each subject.

– "a photo of the nice S^*",
– "a photo of the small S^*",
– "a photo of the weird S^*",
– "a photo of the large S^*",
– "a photo of a cool S^*",
– "a photo of a small S^*"

B.3 Testing Data

For evaluation, we adopt images and editing prompts from DreamBooth [3]. It contains a total of 158 images spanning 30 diverse categories, including dog, cat, robot, boot, etc. Fig. 4 showcases a selection of these image samples. The the complete set of editing prompts for live subjects is detailed below:

– "a S^* in the jungle"
– "a S^* in the snow"
– "a S^* on the beach"
– "a S^* on a cobblestone street"
– "a S^* on top of pink fabric"
– "a S^* on top of a wooden floor"
– "a S^* with a city in the background"
– "a S^* with a mountain in the background"
– "a S^* with a blue house in the background"
– "a S^* on top of a purple rug in a forest"
– "a S^* with a wheat field in the background"
– "a S^* with a tree and autumn leaves in the background"
– "a S^* with the Eiffel Tower in the background"
– "a S^* floating on top of water"
– "a S^* floating in an ocean of milk"
– "a S^* on top of green grass with sunflowers around it"
– "a S^* on top of a mirror"

– "a S^* on top of the sidewalk in a crowded street"
– "a S^* on top of a dirt road"
– "a S^* on top of a white rug"
– "a red S^*"
– "a purple S^*"
– "a shiny S^*"
– "a wet S^*"
– "a cube shaped S^*"

Additionally, we also enumerate the full set of editing prompts for non-live subjects:

– "a S^* in the jungle"
– "a S^* in the snow"
– "a S^* on the beach"
– "a S^* on a cobblestone street"
– "a S^* on top of pink fabric"
– "a S^* on top of a wooden floor"
– "a S^* with a city in the background"
– "a S^* with a mountain in the background"
– "a S^* with a blue house in the background"
– "a S^* on top of a purple rug in a forest"
– "a S^* wearing a red hat"
– "a S^* wearing a santa hat"
– "a S^* wearing a rainbow scarf"
– "a S^* wearing a black top hat and a monocle"
– "a S^* in a chef outfit"
– "a S^* in a firefighter outfit"
– "a S^* in a police outfit"
– "a S^* wearing pink glasses"
– "a S^* wearing a yellow shirt"
– "a S^* in a purple wizard outfit"
– "a red S^*"
– "a purple S^*"
– "a shiny S^*"
– "a wet S^*"
– "a cube shaped $S*$"

References

1. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 3
2. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 4

3. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 5