

Disentangling to Re-couple: Resolving the Similarity-Controllability Paradox in Subject-Driven Text-to-Image Generation

Supplementary Material

6. More Baselines and Benchmark Results

In Tab. 3, we compared against DiptychPrompting[33] and RPO[23]. While RPO achieves high CLIP-T scores, it requires extensive per-instance tuning, making it significantly less feasible for large-scale applications than our approach. Regarding IC-LoRA, its implementation requires task-specific comfyui workflows; we aim to include these results in the final version. Table 4 shows results on DreamBench++. DisCo still outperforms major baselines in subject consistency and prompt adherence across a wider range of subjects.

Table 3. Quantitative results on DreamBench.

Method	CLIP-B-I \uparrow	DINO-I \uparrow	CLIP-B-T \uparrow	ImageReward \uparrow
DreamO	0.899	0.813	0.322	1.186
DreamO-decouple	0.871	0.771	0.312	0.661
UNO	0.899	0.827	0.311	0.854
UNO-decouple	0.891	0.828	0.303	0.359
Diptych Prompting	0.864	0.767	0.319	1.136
RPO	0.852	0.725	0.338	1.164
DisCo (ours)	0.928	0.903	<u>0.329</u>	1.339

Table 4. Quantitative results on DreamBench++.

Method	CLIP-B-I \uparrow	DINO-I \uparrow	CLIP-B-T \uparrow	ImageReward \uparrow
OminiControl	0.757	0.538	0.335	<u>1.173</u>
ACE++	0.787	0.552	0.324	0.762
DreamO	<u>0.795</u>	<u>0.589</u>	<u>0.336</u>	1.044
UNO	0.782	0.549	0.321	0.693
FLUX.1 Kontext [dev]	0.790	0.572	0.334	1.118
DisCo (ours)	0.801	0.610	0.339	1.291

7. Qualitative Results with General Text-to-image Models

Figure 8 provides a qualitative comparison of DisCo against several leading general image editing models: Nano Banana, Qwen-Image-Edit[42], and Seedream 4.0[31]. Our method achieves a superior balance of subject similarity and text controllability, showing performance that is comparable to or even surpasses these state-of-the-art methods.

However, other baselines exhibit specific drawbacks. Nano Banana occasionally compromises image authenticity; for instance, in the `clock` example, it spliced half of the reference image with the Eiffel Tower, generating a fabricated photograph. Qwen-Image-Edit struggles with maintaining correct proportions, as seen in the `dog` example where the scale appears disproportionate. Furthermore, Seedream 4.0 demonstrates relatively poor subject fidelity. Among all methods, DisCo successfully mitigates

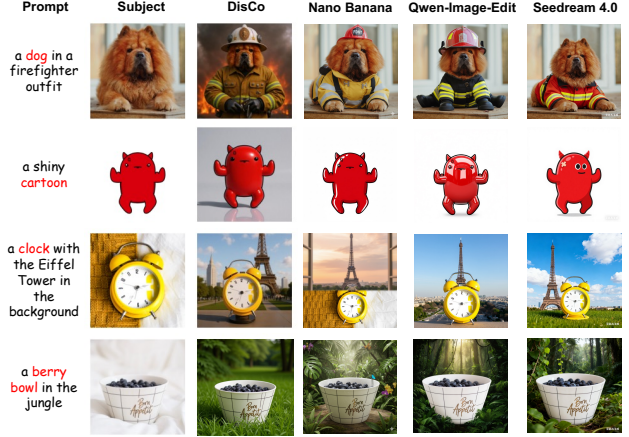


Figure 8. Qualitative results with leading general image editing models.

these specific weaknesses, achieving the most balanced performance across subject consistency, seamless integration, and overall image quality.

You're writing a image-editing instruction to edit this image of the subject "subject" in three ways:

1. Change the identity attributes of the subject (e.g., color, texture, shape) to make it slightly different.
2. Introduce minor distortions or artifacts to the image that reduce its visual clarity or fidelity
3. Change the way the subject interacts with the environment or to create a fake or unnatural scene.

Choose one of these strategies and write a concise instruction (1 to 2 sentences) that clearly describes how to make the image worse.

Return in this json format:

```
{
  "edit_instruction": the instruction that you wrote,
}
```

Figure 9. The prompt used to generate editing instructions.

8. Details of Reward Modeling

Evaluating subject-driven text-to-image tasks requires a comprehensive understanding of both images and text instructions. In this work, we employ a Vision Language

Model (VLM) for reward modeling, capitalizing on its multimodal understanding capabilities. While a straightforward approach would be pointwise evaluation (i.e., scoring a single image), such methods are often sensitive to minor prompt variations and can produce noisy reward signals. Our reward modeling, therefore, builds on a pairwise comparison framework within the Group Relative Policy Optimization (GRPO) stage.

The core of this framework is the computation of the preference probability $P_\phi(x^i \succ x^j \mid c_I, c_T)$, which indicates that the generated image x^i is preferred over x^j . c_I and c_T are the reference image and text instruction, respectively, and ϕ is a preference predictor parameterized by a VLM. We compute this probability by presenting both images, x^i and x^j , to a trained VLM ϕ and prompting it to select the superior one given c_I and c_T . The normalized probability of the output token corresponding to the choice “ x^i is better” is then used as the value for $P_\phi(x^i \succ x^j \mid c_I, c_T)$.

With this mechanism, the reward for a sample x_0^i within a group of size G during GRPO training is defined as the sum of its win probabilities against all other samples in the group:

$$R_i = \sum_{\substack{1 \leq j \leq G \\ j \neq i}} P_\phi(x_0^i \succ x_0^j \mid c_I, c_T) \quad (7)$$

To get more robust preference predictions, we build a dataset of preference pairs and train the VLM ϕ . Instead of relying on manual annotation, we generate a synthetic dataset using the image generation model itself. For a given “positive” example (c_I, c_T, x_0) from the dataset, a VLM first crafts a “negative” editing instruction \tilde{c}_T . This instruction intentionally introduces flaws, such as altering the subject’s key attributes, creating inconsistencies with the original prompt c_T , or producing unnatural compositions. Subsequently, we generate a negative sample \tilde{x}_0 by applying the editing instruction \tilde{c}_T to the original image x_0 using the image generation model. This process yields a preference pair $(x_0 \succ \tilde{x}_0)$, where the original image is preferred over the corrupted one.

The VLM ϕ is then trained by minimizing the negative log-likelihood of these preferences:

$$\mathcal{L}_\phi = \mathbb{E}_{(x_0, \tilde{x}_0, c_I, c_T) \sim \mathcal{D}} [-\log P_\phi(x_0 \succ \tilde{x}_0 \mid c_I, c_T)], \quad (8)$$

where \mathcal{D} is our synthetically generated dataset of preference tuples.

The prompts used to generate the negative editing instructions \tilde{c}_T and to elicit pairwise judgements from ϕ are provided in Fig. 9 and Fig. 10, respectively.

9. Prompt for VLM in the TVD Module

We utilize Qwen2.5-VL-72B-Instruct as the VLM for textual-visual decoupling. The model is assigned two pri-

You're labeling some preference data to train a subject-driven image generating model, where

1. The first given image is the original one (reference image), and the main subject “{subject}” is marked by a red bounding box.
2. The second and third images are the generated results
3. The generation is performed under this text instruction: “{instruction}”.

You should choose one of the two generated results based on these criterion:

1. Composition plausibility: The subject should be seamlessly and naturally integrated into the new context, with coherent lighting, perspective, and style.
2. Subject Consistency: the main subject should not deviate too much from that in the reference image, except indicated by the text instruction or changes in lighting, angle, size, position, etc.
3. Absence of Irrelevant Artifacts: The generated image should not retain any elements from the reference image that are unrelated to the subject (e.g., background remnants, unrelated objects, or visual clutter).
4. Overall image quality: The image should be free from visual artifacts, distortions, or unnatural features that detract from its realism and aesthetic appeal.

Return a single letter for the better image: "A" if the second image is better, or "B" if the third image is better.

Figure 10. The prompt used to train the reward model.

mary tasks: (1) identifying the main subject by jointly analyzing the reference image and the input text prompt, and (2) rewriting the prompt such that the identified subject is replaced with a generic pronoun. The specific prompt used to guide VLM in this process is provided in Fig. 11.

You are given:

1. A reference image.
2. A text prompt for a subject-driven image generation task, which involves the main subject inside the reference image.

Your task:

1. Identify the main subject in each image. Just the main subject, no description of the background.

2. Rewrite the prompt following these steps:

(1) Replace subject name with a single, generic pronoun or determiner, such as it, he/she, they, 'this item'.

(2) Keep other descriptions in the prompt intact, such as the background, the transformations of the subject, or other elements.

(3) (Optional) Paraphrase the sentence to accommodate the pronoun. This should only be for grammar purposes.

NOTE: You can NOT describe anything about the reference image in your rewritten prompt.

Return in this json format:

```
{  
  "subject": the main subject in the reference image,  
  "prompt": the prompt that you rewrote  
}
```

Figure 11. The prompt used in the TVD module.