Scone: Bridging Composition and Distinction in Subject-Driven Image Generation via Unified Understanding-Generation Modeling

Supplementary Material

Contents

1. Additional details of motivation

The similarity visualizations of instruction and image tokens in Figs.1(b) and 2(a) of the main paper are based on our base model, BAGEL [1]. To better highlight high-similarity regions, we retain the top 50% and generate masked images. We group layers into four sets (0-7, 8-15, 16-23, 24-27) based on the layer function analysis from [6]. In Fig. 2, we further show representative similarity and masked images for each group.

Observation 1 (Comparison between understanding and generation experts) The understanding expert captures more distinct semantic information from the image than the generation expert, as its image-token hidden states exhibit higher similarity to the instruction-token hidden states. It attends more strongly to instruction-relevant regions, such as candidate subjects.

Observation 2 (Comparison across layers of the understanding expert) Although similarity remains high at layers 16 and 24, region discrimination is strongest at layer 8, which provides more distinctive semantic cues for generation guidance. We therefore choose layer 8 to provide the semantic mask, and apply it to the later semantically discriminative layers 9–15.

2. Additional details of training data

2.1. Synthesized data for data pool

As described in Sec.5.1 of the main paper, we synthesize 15K samples with 3-4 input images to fill gaps in the data pool and improve the composition capability of Scone. Examples are shown in Figs. 4 and 5.

Prompt for distinction scoring

```
### Given
- A **subject description**.
- The **first image** is the reference image.
- The **second image** is the target image.

### Task
Determine whether the described subject from the reference image **appears**
in the target image.

1. **Identify** the subject in the reference image based on the given description.
2. **Judge presence** in the target image:
- Focus strictly on **presence**, not on appearance similarity or instruction
compliance.
- Assign **1** if the subject is clearly identifiable in the target image.
- Assign **0** if the subject is not identifiable.
...
### IMPORTANT: Your response must be either 0 or 1.
...
```

Figure 1. Prompt for distinction scoring in SconeEval. It determines whether the described reference subject *appears* in the target image.

2.2. Data filtering for refined single-candidate data

As described in Sec.5.1 of the main paper, refined single-candidate samples are filtered by scoring subject consistency and instruction following with the VLM model Qwen3-VL-30B-A3B-Instruct. Key prompt contents are shown in Fig. 6(a), with emphasis on facial identity, text, and quantity. Each sample is scored from 0 to 4, and only those with a score of 4 are retained, as shown in Fig. 6(b).

2.3. Details of multi-candidate data

Single-subject data Multi-candidate single-subject data are derived from single-candidate multi-subject data by reversing the reference and target images, so that the original reference becomes the target and the original target becomes the reference. This avoids the cost of generating new images. For instruction construction, Qwen3-30B-A3B-Instruct-2507 [3] identifies subjects, provides distinctive descriptions, and generates instructions based on the prompt in Fig. 7(a). The final dataset contains 2 case types, each with cross-category and intra-category candidate subjects. Examples are shown in Fig. 7(b).

Multi-subject data Multi-subject data are constructed from single-candidate multi-subject data by editing a subset of the reference images. Specifically, we use GPT-4o [2] to generate prompts for subjects from different categories, and then add at least one subject to the reference images using Qwen-Image-Edit-2509 [4]. The instruction

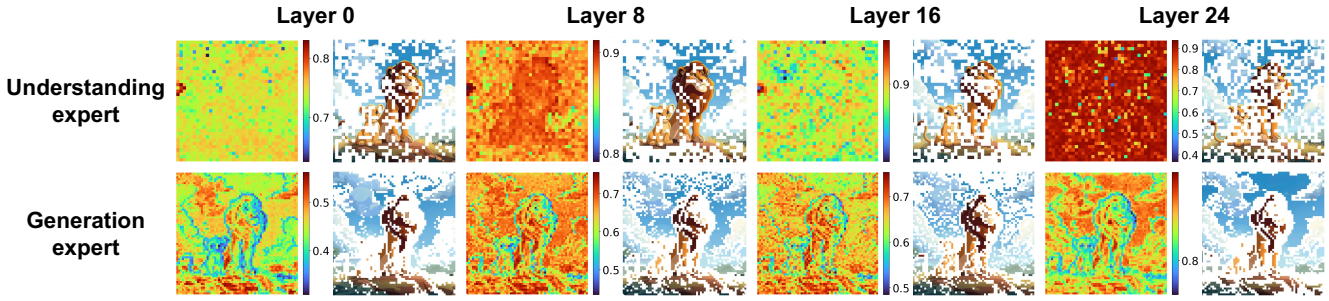

Figure 2. Representative similarity and masked images for each layer group. Similarity visualizations of instruction and image token hidden states from understanding and generation experts are based on our base model, BAGEL [1]. Masked images retain the top 50% of regions for clearer observation.

construction consists of two steps: *(1) subject identification* and *(2) subject replacement*. Subject identification follows the procedure described in Sec.4.2 of the main paper, with additional details provided in Sec. 3. Subject replacement uses Qwen3-30B-A3B-Instruct-2507 [3] and the prompt in Fig. 8(a) to replace the original subject description with the distinct description obtained in Step 1. The final dataset contains 5 case types, each including both cross-category and intra-category candidate subjects in the reference images. Examples are shown in Fig. 8(b).

3. Two-step decoupling instruction construction in SconeEval

As described in Sec.4.2 of the main paper, we adopt a two-step decoupling strategy that separates visual understanding from instruction generation, improving instruction stability and quality. As shown in Fig. 9, direct instruction construction often produces unusable results, such as incorrect image indices, ambiguous target subjects, and unrelated subjects. Our strategy first uses the vision-language model Qwen3-VL-30B-A3B-Instruct [3] to identify the target subject and generate a distinct description from the raw single-candidate and edited multi-candidate reference images, with prompt in Fig. 10(a). It then uses the language model Qwen3-30B-A3B-Instruct-2507 [3] to generate instructions solely from the subject descriptions, with prompt in Fig. 10(b).

4. Limitation and future work

Our Scone still shares a common limitation with existing methods: unrealistic interactions. As shown in Fig. 3, the generated dog passes through the chair, violating physical laws. This issue appears in both our Scone and OmniGen2 [5]. Future work will also explore more efficient mechanisms to reduce redundant image tokens for scalable generation in complex scenarios.

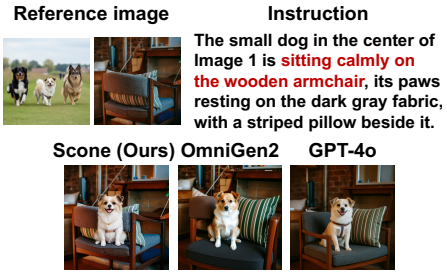

Figure 3. Limitation of our Scone.

References

[1] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 1, 2

[2] OpenAI. Hello gpt-4o, 2025. 1

[3] Qwen Team. Qwen3 technical report, 2025. 1, 2

[4] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 1

[5] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2

[6] Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19781–19791, 2025. 1

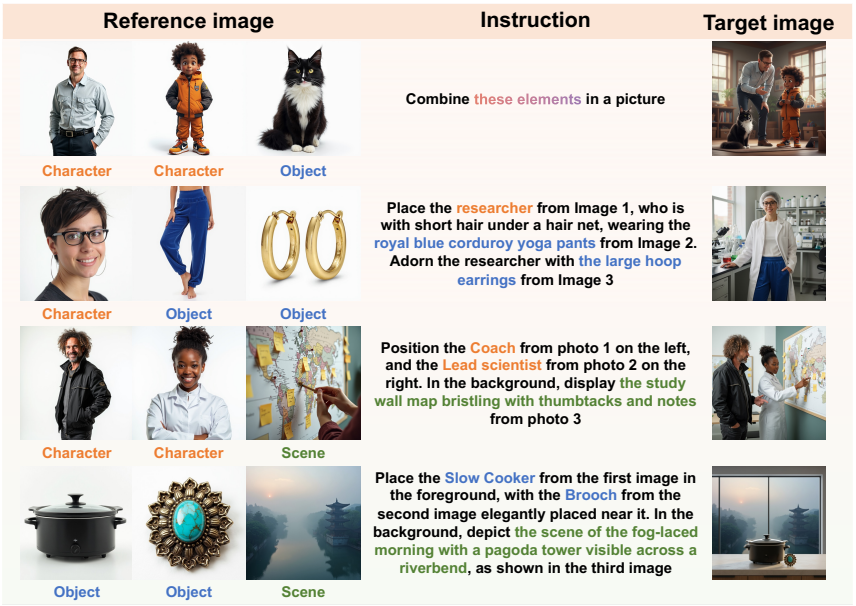

Figure 4. Examples of synthesized data with 3 input images. These cover 4 case types, including character-character/object interactions, characters with multiple objects, characters in scenes, and objects placed in scenes.

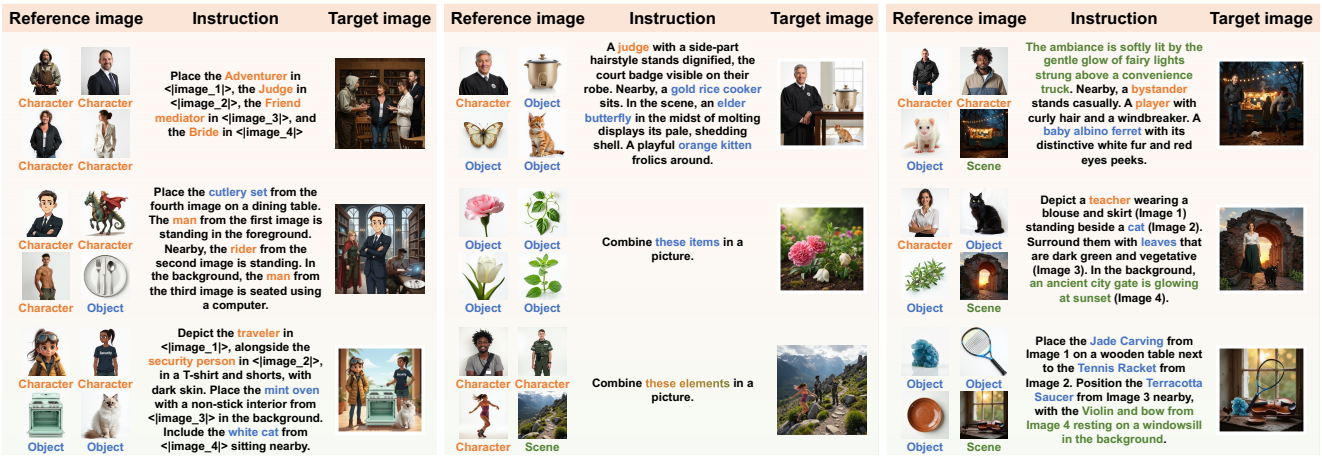

Figure 5. Examples of synthesized data with 4 input images. These cover 9 case types, including combinations of characters, objects, and scenes, as well as their interactions and mixed compositions.

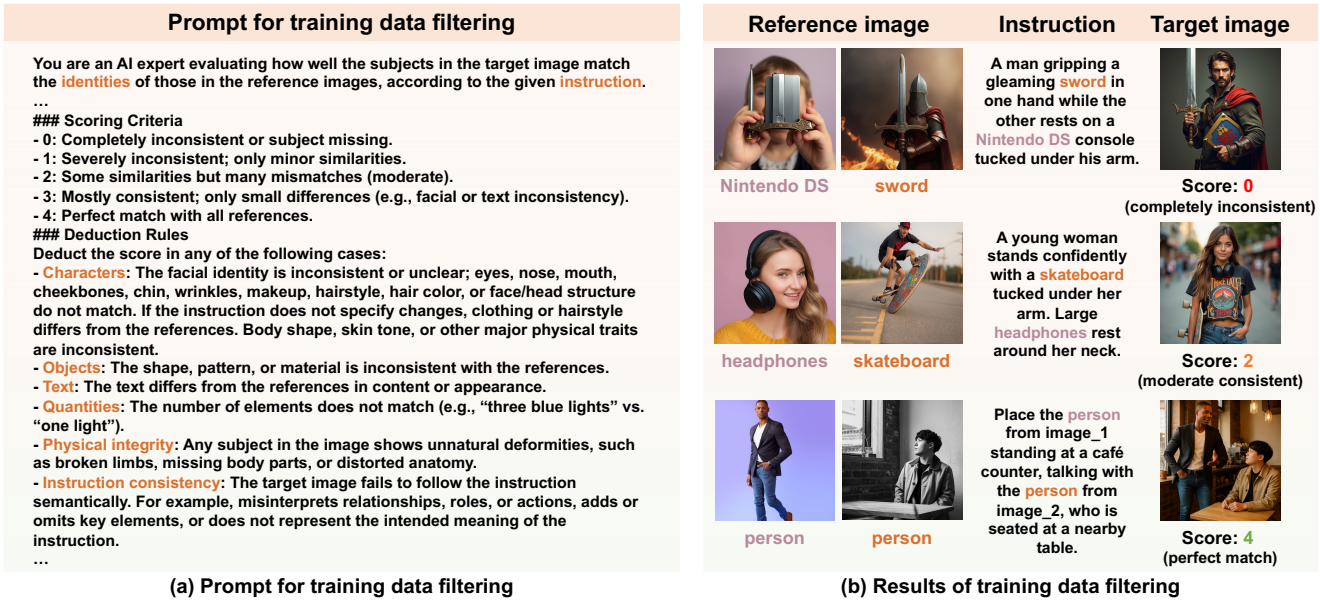

Figure 6. Data filtering for refined single-candidate data. (a) Prompt for filtering. Key prompt components are shown. **(b) Filtering results.** Samples are scored from 0 to 4, and only those with a score of 4 are retained.

Figure 7. Multi-candidate single-subject data construction. (a) Prompt for instruction construction. The prompt instructs the vision-language model to identify subjects, provide distinct descriptions, and generate instructions. **(b) Example demonstration.** It includes 2 case types, Character and Object, each with cross-category and intra-category candidate subjects in the reference images.

Figure 8. Multi-candidate multi-subject data construction. (a) Prompts for subject replacement. The prompts guide the language model to replace the original subject description with a new distinct description for the edited multi-candidate reference images. **(b) Example demonstration.** It includes 5 case types: Character+Character, Character+Object, Object+Object, Character+Scene, and Object+Scene, each with cross-category and intra-category candidate subjects in the reference images.

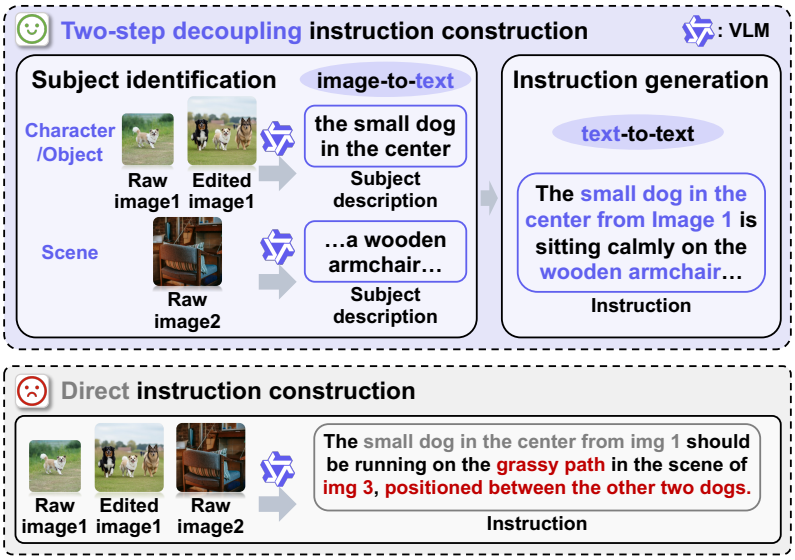

Figure 9. Comparison of two instruction construction strategies. The two-step decoupling strategy separates image-to-text and text-to-text generation, reducing cross-image interference and avoiding errors in the direct strategy, such as incorrect image indices, ambiguous target subjects, and unrelated subjects.

Prompt for subject identification

For Character/Object

You are an AI assistant skilled at identifying a target subject in an image based on a reference image.

Task
You are given two images.
- The first image (reference image) contains the target subject.
- The second image (image to identify) contains multiple subjects.

Your goal is to **identify and describe only the subject in the second image** that corresponds to the reference.
Provide an **unambiguous, concise phrase** that uniquely distinguishes the target subject **within the second image**.

Use cues such as:
- **attribute**: …
- **position**: …
- **size**: …
…

For Scene

Task
Given a scene image, describe the scene clearly and concisely.
1. Provide a one-sentence summary of the overall scene, including whether it's indoors or outdoors and the general setting (e.g., city street, park, living room).
2. List the main and most noticeable objects, such as furniture, people, animals.
…

(a) Prompt for subject identification

Prompt for instruction generation

For samples w/o Scene

You are an AI assistant skilled in composing subject-driven image generation instructions.

Task
Given a list of subject referring descriptions (each corresponding to an input image index), write a single, coherent instruction that places all subjects in a new state or scene.
- If there is only one description, describe that subject in a new state or scene.
- If there are multiple descriptions, describe how the subjects appear or interact together naturally in a new scene.
…

For samples w/ Scene

You are an AI assistant skilled in composing subject-driven image generation instructions.

Task
Given a list of descriptions (each corresponding to an input image index), write a single, coherent instruction that places all subjects in the specified scene.
- The former description(s) refer to subjects.
- The last description refers to the scene, showing the background or environment where the subjects should appear.

Construct a natural, realistic instruction describing how the subjects appear or interact within the scene (e.g., the bird is standing on the table from the scene).
…

(b) Prompt for instruction generation

Figure 10. Prompts for instruction construction in SconeEval. (a) Prompt for subject identification. For Character or Object images, provide a concise subject description; for Scene images, describe the setting and key objects. **(b) Prompt for instruction generation.** Generate instructions from the subject descriptions, emphasizing subject-subject and subject-scene interactions.