Supplemental Material: DreamOmni2: Multimodal Instruction-based Editing and Generation

Bin Xia[1,4] Bohao Peng[1] Yuechen Zhang[1] Junjia Huang[4],
Jiyang Liu[4] Jingyao Li[1] Haoru Tan[3] Sitong Wu[1] Chengyao Wang[1],
Yitong Wang[4] Xinglong Wu[4] Bei Yu[1] and Jiaya Jia[2]
[1]CUHK [2]HKUST [3]HKU [4]ByteDance Inc

https://github.com/dvlab-research/DreamOmni2

The overview of the supplementary materials:

(1) We discussed the differences between the current unified generation and understanding models. (Sec. 1).

(2) We discussed the failure cases of VLM evaluation. (Sec. 2).

(3) We provide more details on the joint training of VLM and Kontext (Sec. 3).

(4) We compared the CLIP and DINO metrics in the multimodal instruction-based editing task to measure the consistency of the editing results (Sec. 4).

(5) We provide the system prompt for model evaluation and additional evaluation details (Sec. 5).

(6) We present more visual comparisons of DreamOmni2 and other competitive methods on Multimodal Instruction-based Editing and Generation (Sec. 6).

(7) We provide more details on the DreamOmni2 benchmark (Sec. 7).

(8) We present a large number of DreamOmni2 visual results on Multimodal Instruction-based Editing (Sec. 8).

(9) We present a large number of DreamOmni2 visual results on Multimodal Instruction-based Generation (Sec. 9).

1. Discusiion about understanding&generation models.

(1) We have compared the most recent unified understanding&generation model, OmniGen2, in paper Tab. 1 and Fig. 5. DreamOmni2 outperforms them. **(2)** Editing requires maintaining consistency in non-edited regions and supporting both abstract attribution and concrete objects editing. Previous methods focus on subject-driven generation and do not enforce non-edited consistency; without an explicit training data creation scheme, they perform poorly for multimodal editing.

2. Failure Cases of VLM Evaluation.

(1) For DreamOmni2, VLM may incorrectly flag some insertion tasks due to improper scaling, even though the outputs are visually acceptable. **(2)** For GPT-4o, its non-edited region consistency is poor, but VLM fails to detect changes.

3. Joint Training of VLM and Kontext

For the joint training of VLM and Kontext, we use a two-phase training approach. The training data is shown in Fig. 1. Specifically, we prepare predefined standard instructions and user instructions. In the first phase, we train the VLM to translate user instructions into predefined standard instructions. In the second phase, we use images and predefined standard instructions as inputs to Kontext, training the model to generate the target image.

4. Traditional Metric

As shown in Table 1, we compared the CLIP-I and DINO-I between the edited image and the source image in the multimodal instruction-based editing task to further evaluate the pixel consistency of the edited images. It can be observed that the consistency of our DreamOmni2 is significantly better than open-source models, and even approaches that of commercial models.

5. Model Evaluation

As shown in Tabs. 2 and 3, we provide the system prompts used for evaluating multimodal instruction-based editing and generation with Doubao [2] and Gemini [3] in the paper.

For the human evaluation in the Tab. 2 and 3 of the paper, we use the same evaluation criteria as described in the system prompt above. Specifically, for **multimodal instruction-based editing**, we assess the following four aspects:

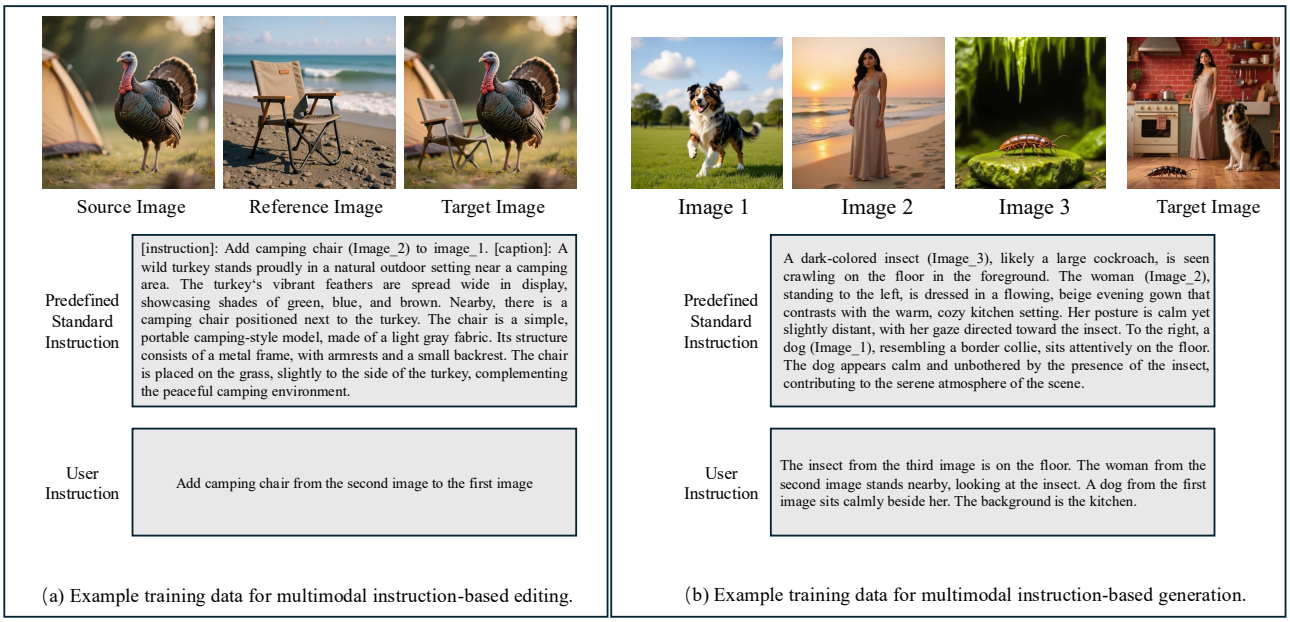


Figure 1. Example training data for multimodal instruction-based editing and generation.

Table 1. Comparison of consistency under CLIP and DINO.

Method	Concrete Object		Abstarct Attribution	
	CLIP-I↑	DINO↑	CLIP-I↑	DINO↑
GPT-4o [6]	0.7749	0.6135	0.8265	0.7065
Nano-banana [4]	0.7903	0.6529	0.8934	0.8348
UNO [9]	0.6171	0.4334	0.6375	0.4343
DreamO [5]	0.6802	0.5526	0.7357	0.5873
Omnigen2 [8]	0.7453	0.5625	0.8061	0.7049
Qwen-Image-Edit [7]	0.6990	0.5014	0.7152	0.5625
Kontext [1]	0.7279	0.5544	0.7646	0.5866
Qwen-Image-Edit-2509 [7]	0.7382	0.5421	0.8520	0.7332
DreamOmni2 (Ours)	**0.7564**	**0.5879**	**0.8621**	**0.7671**

- **Consistency of Non-Edited Regions**: Evaluate whether the regions or objects in the source image that were not edited (as per the instructions) maintain consistency with the target image. Specifically, compare the appearance and positioning of these objects in both images to ensure they remain identical. Also, assess whether there are any noticeable color discrepancies, such as yellowing, whitening, sharpening, or blurring, that may indicate inconsistencies.
- **Consistency of Edited Elements**: Examine the objects or elements in the target image that were edited according to the instructions. Compare these edited areas with the reference image to ensure that the requested changes were accurately applied, maintaining visual coherence.
- **Instruction Accuracy**: Verify whether the editing instruction has been accurately followed. For example, if the instruction specifies placing a woman from the reference image onto a sofa in the source image, confirm that the woman not only matches the reference in terms of appearance but is also properly positioned on the sofa, as instructed.
- **Quality of the Editing Result**: Assess the overall quality of the edits in the target image. Compare the edited objects with their counterparts in the reference image to determine if the changes appear natural. Ensure that the edited elements blend seamlessly with the surrounding environment and that no parts of the object appear misplaced, incomplete, or poorly integrated.

We will ask five evaluators to assess each multimodal instruction-based editing case based on the four criteria above. If any criterion is not met, the case will be marked as a failure; otherwise, it will be considered a successful edit. A case is deemed successful if at least three evaluators judge it as such; otherwise, it will be classified as a failure.

For **multimodal instruction-based generation**, we assess the following four aspects:

- **Consistency**: Assess whether the elements specified in the instructions (such as people, objects, or abstract attributes) are accurately reflected in the target image, ensuring they align with their counterparts in the reference image.
- **Instruction Adherence**: Evaluate whether the target image has been generated according to the given instructions. Confirm that all requested changes and features have been correctly implemented.
- **Visual Integrity**: Check for any noticeable visual issues

such as unrealistic blending, unnatural proportions, or mismatched lighting and shadows that might compromise the overall look of the image.

- **Generation Quality**: Ensure that the reference objects or attributes have not been overly copied or duplicated, which could lead to unnatural repetition or inconsistencies in the image. The final result should feel cohesive and authentic, rather than forced or artificially created.

Similarly, we will have five evaluators assess each multimodal instruction-based edit based on the four criteria above. If any criterion is not met, the case is marked as a failure. A case is considered successful if at least three evaluators agree; otherwise, it is classified as a failure.

6. More Comparisons on Multimodal Instruction-based Editing and Generation

We provide more visual results comparisons on multimodal instruction-based editing and generation in Fig. 4 and Fig. 5, respectively.

7. DreamOmni2 Benchmark

Our DreamOmni2 benchmark includes 205 multimodal instruction-based editing test cases and 114 instruction-based generation test cases. Visualizations of the editing and generation test cases are shown in Fig. 4 and Fig. 5, respectively. The benchmark covers a wide range of test cases, with input reference images ranging from one to five, and encompasses diverse local and global attributes, as well as concrete objects. The DreamOmni2 Benchmark will be released.

8. More Multimodal Instruction-based Editing Cases

As shown in Fig. 6, Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig. 17, Fig. 18, and Fig. 19, we present additional visual cases of DreamOmni2 on the multimodal instruction-based editing task.

9. More Multimodal Instruction-based Generation Cases

As shown in Fig. 20, Fig. 21, Fig. 22, Fig. 23, Fig. 24, Fig. 25, Fig. 26, Fig. 27, and Fig. 28, we present additional visual cases of DreamOmni2 on the multimodal instruction-based generation task.

References

[1] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 2

[2] ByteDance. Doubao. https://www.doubao.com/, 2025. 1

[3] Google. Gemini. https://deepmind.google/models/gemini/, 2025. 1

[4] Google. Nano banana. https://aistudio.google.com/models/gemini-2-5-flash-image, 2025. 2

[5] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 2

[6] OpenAI. Gpt-4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. 2

[7] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2

[8] 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

[9] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *ICCV*, 2025. 2

You are a professional image editing evaluator. I will provide you with a source image, a reference image, a target image, and an editing instruction. Your task is to evaluate the target image based on the following criteria: 1. Consistency of Non-Edited Regions: Assess whether the regions or objects in the source image that were not edited (as per the instructions) are consistent with the target image. Specifically, compare whether the appearance and position of each object in the non-edited areas of the source and target images remain strictly consistent, and evaluate if there are any significant color discrepancies, such as yellowing, whitening, sharpening, blurring, etc. 2. Consistency of Edited Elements: Evaluate the consistency of the objects or elements in the target image that were edited based on the instruction. Compare the reference image and target image to ensure the requested edits were applied properly. 3. Instruction Accuracy: Check if the editing instruction has been accurately executed. For instance, if the instruction asks to place a woman from the reference image onto a sofa in the source image, verify if the woman in the target image not only looks similar to the reference image but also correctly sits on the sofa as instructed. 4. Quality of the Editing Result: Assess the overall quality of the edits in the target image. Compare the edited objects between the reference image and the target image to determine if the edits appear unnatural, such as if the person or object has been copied from the reference image into the target image in a way that doesn't blend well. Also, evaluate whether any parts of the object are missing or if the object doesn't seamlessly integrate with the surrounding environment.
Output the result in JSON format as: "judge": True/False, "reason": "xxx". Please analyze the requirements step by step according to the four points. Only return True if all points are satisfied. For example, if the source image shows a girl dancing and the instruction asks for the boy in the target image to have the same pose as the girl in the reference image, your task is to: Check if the parts of the source image that weren't mentioned in the instruction (like the background or other elements) remain consistent in the target image. Ensure that the boy's pose in the target image matches the girl's pose in the reference image. If everything is consistent, the editing is good, and the requested changes have been made correctly, return True. If there are noticeable discrepancies or significant issues, return False, and provide a clear reason for your judgment.

Table 2. The system prompt for the VLM (Doubao and Gemini) to assess multimodal instruction-based editing.

You are a professional reference image generation evaluator. I will provide you with reference images, the final target image, and an instruction for generating the target image based on the reference image. Your task is to evaluate the target image based on the following criteria:
1. Consistency: Check whether the elements requested in the instruction, which are derived from the reference image (such as people, objects, or abstract attributes), are consistent with the corresponding elements in the target image. 2. Instruction Adherence: Verify whether the target image has been generated accurately according to the given instruction. Ensure that the changes and features requested in the instruction are correctly implemented in the target image. 3. Visual Integrity: Assess whether there are any noticeable issues of visual breakdown or distortion in the target image (such as unrealistic blending, unnatural proportions, or mismatched lighting and shadows). 4. Generation Quality: Evaluate whether the reference objects or attributes have been directly copied and pasted, which could lead to inconsistencies such as missing parts, or unnatural duplication, or result in elements in the target image that are too similar to the reference, making the final image feel forced or unnatural. Output the result in JSON format as: "judge": True/False, "reason": "xxx"
Please analyze according to the three points step by step. Only return True if all points are satisfied. For example, if the instruction asks to place a person from the reference image into a specific position in the target image, you should check whether the person's appearance, pose, and position in the target image align with the reference image, and whether any other requested details are properly followed. Additionally, assess if there are any visible flaws like disproportionate figures, inconsistent lighting, or mismatched style. If everything aligns correctly and no major issues are present, return True. If there are discrepancies or noticeable flaws, return False and provide a clear explanation for your judgment.

Table 3. The system prompt for the VLM (Doubao and Gemini) to assess multimodal instruction-based generation.

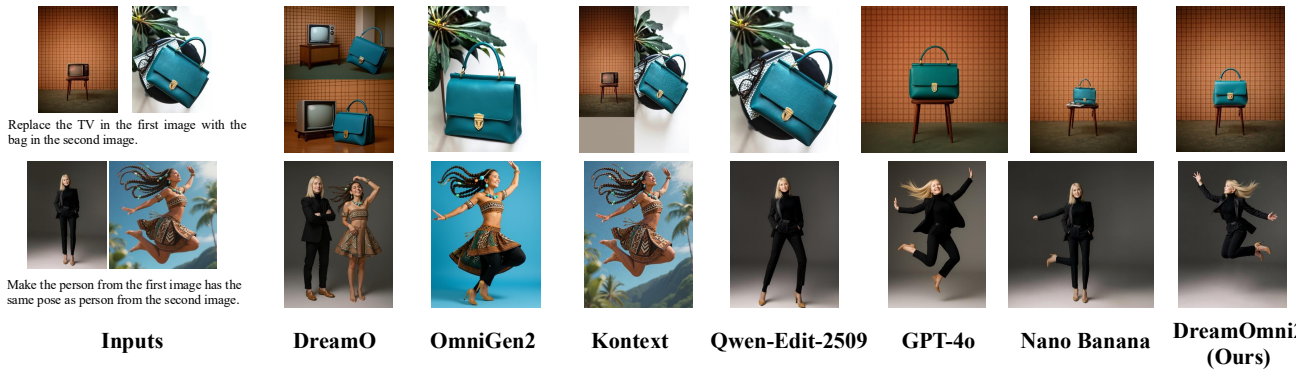


Figure 2. More visual comparison of multimodal instruction-based editing. Compared to other competitive methods and even closed-source commercial models (GPT-4o and Nano Banana), DreamOmni2 shows more accurate editing results and better consistency.

Figure 3. More visual comparisons of multimodal instruction-based generation. Our DreamOmni2 significantly outperforms current open-source models and achieves generation results comparable to closed-source commercial models (GPT-4 and Nano Banana).

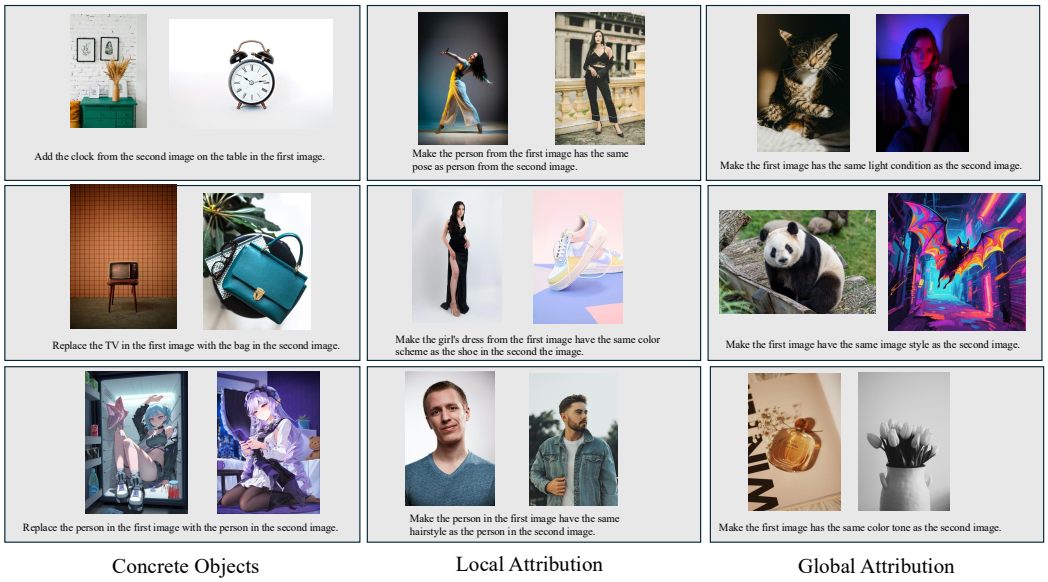


Figure 4. Examples of multimodal instruction-based editing in DreamOmni2 benchmark.

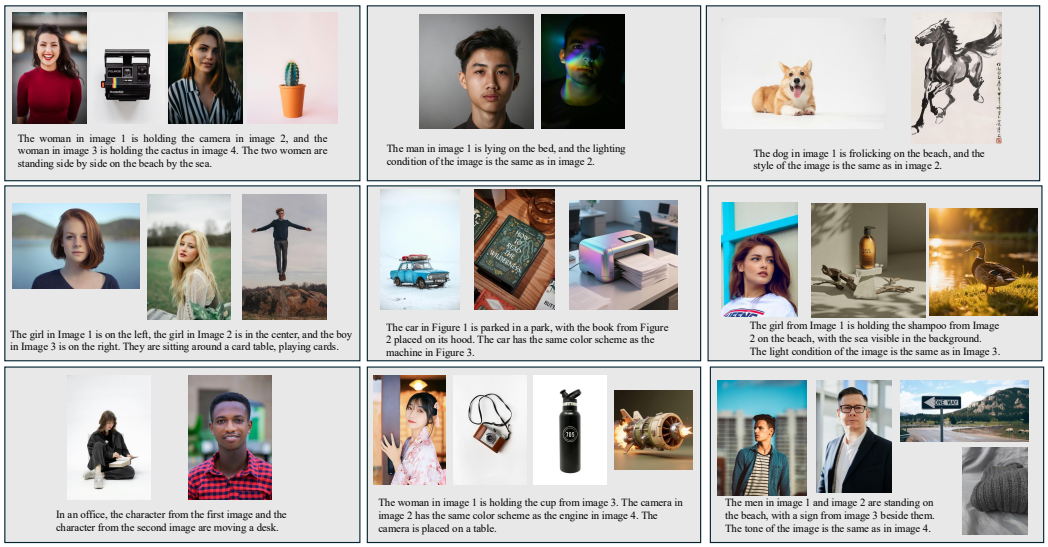


Figure 5. Examples of multimodal instruction-based generation in DreamOmni2 benchmark.

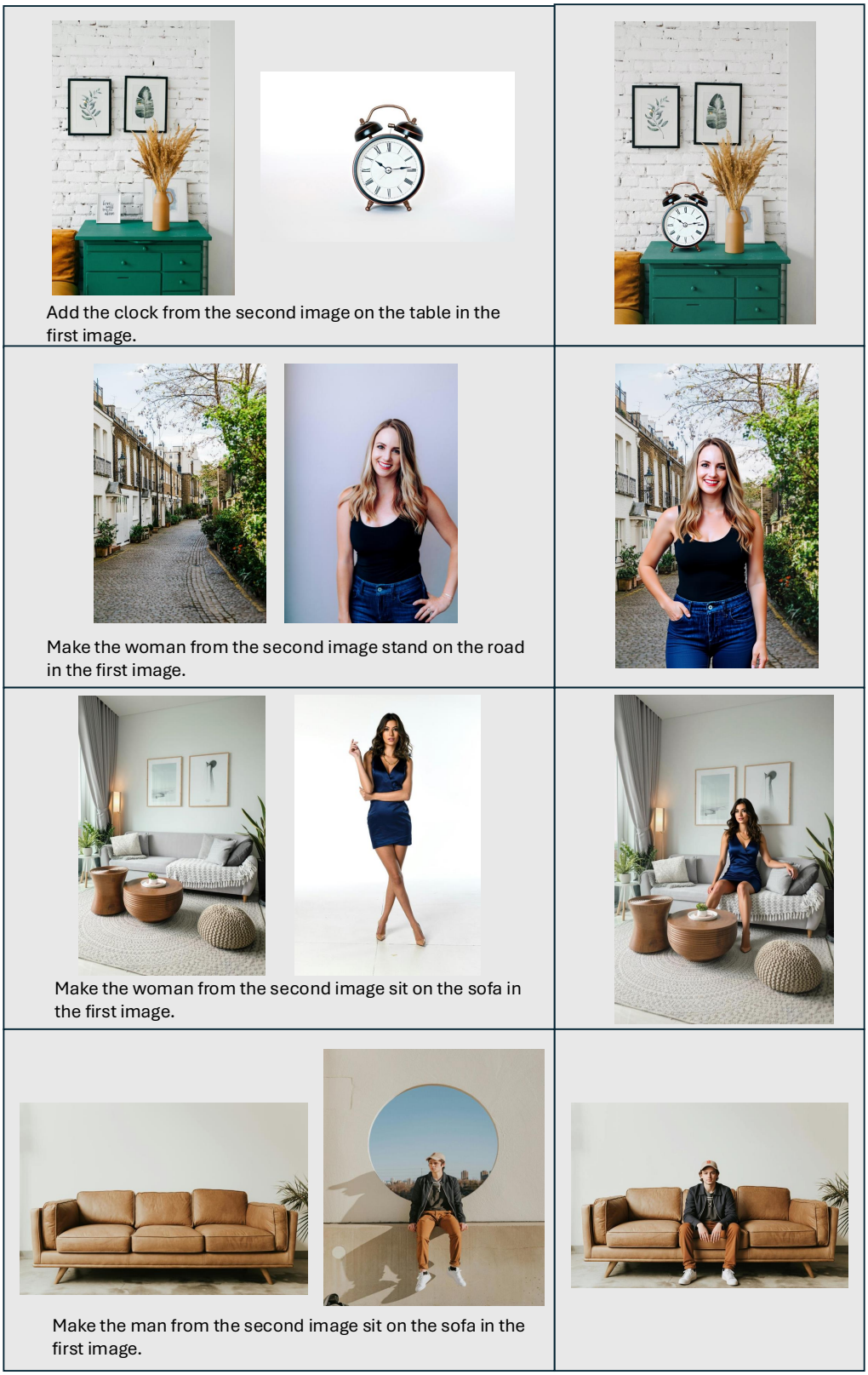


Figure 6. Multimodal instruction-based editing cases of DreamOmni2.

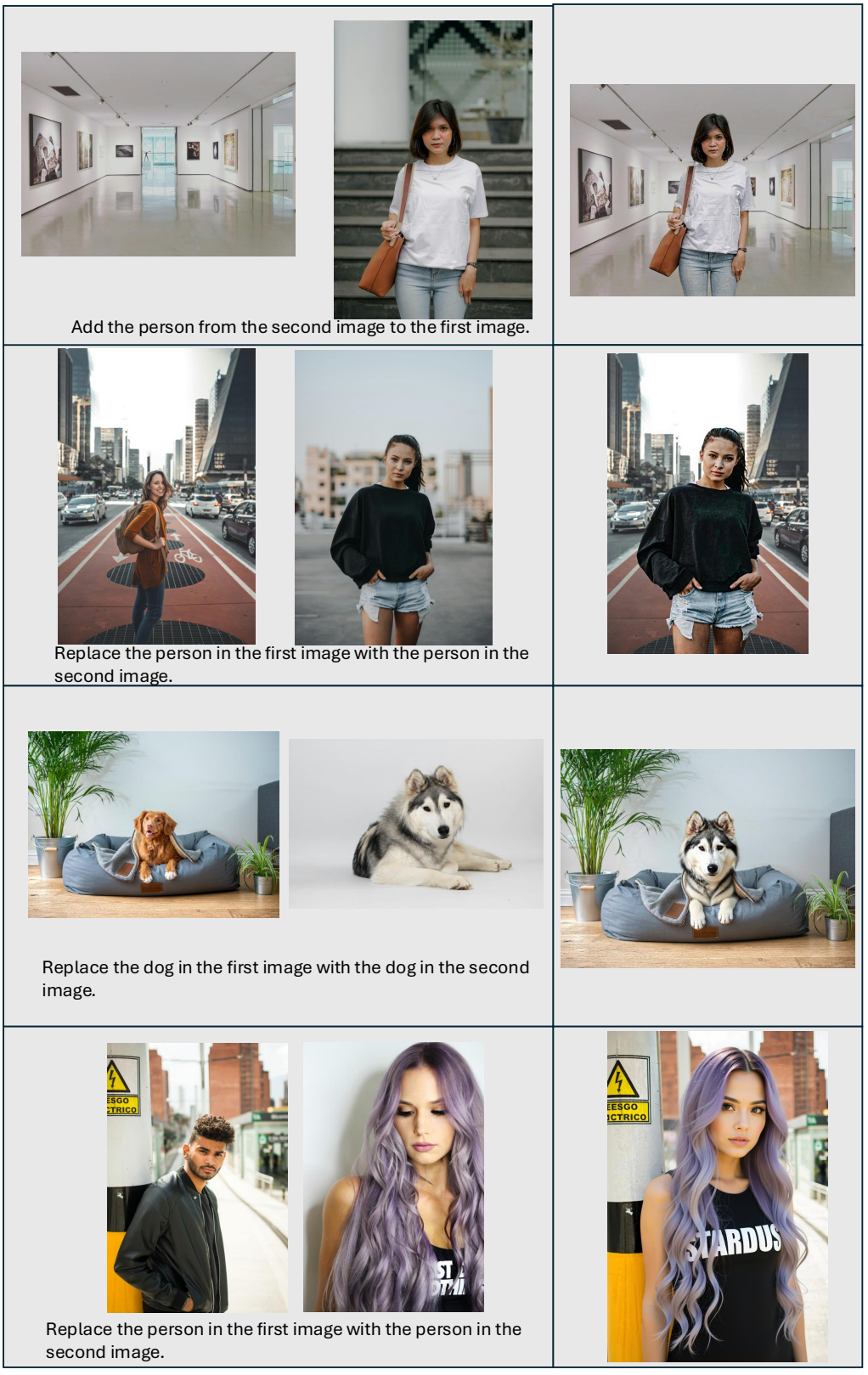


Figure 7. Multimodal instruction-based editing cases of DreamOmni2.

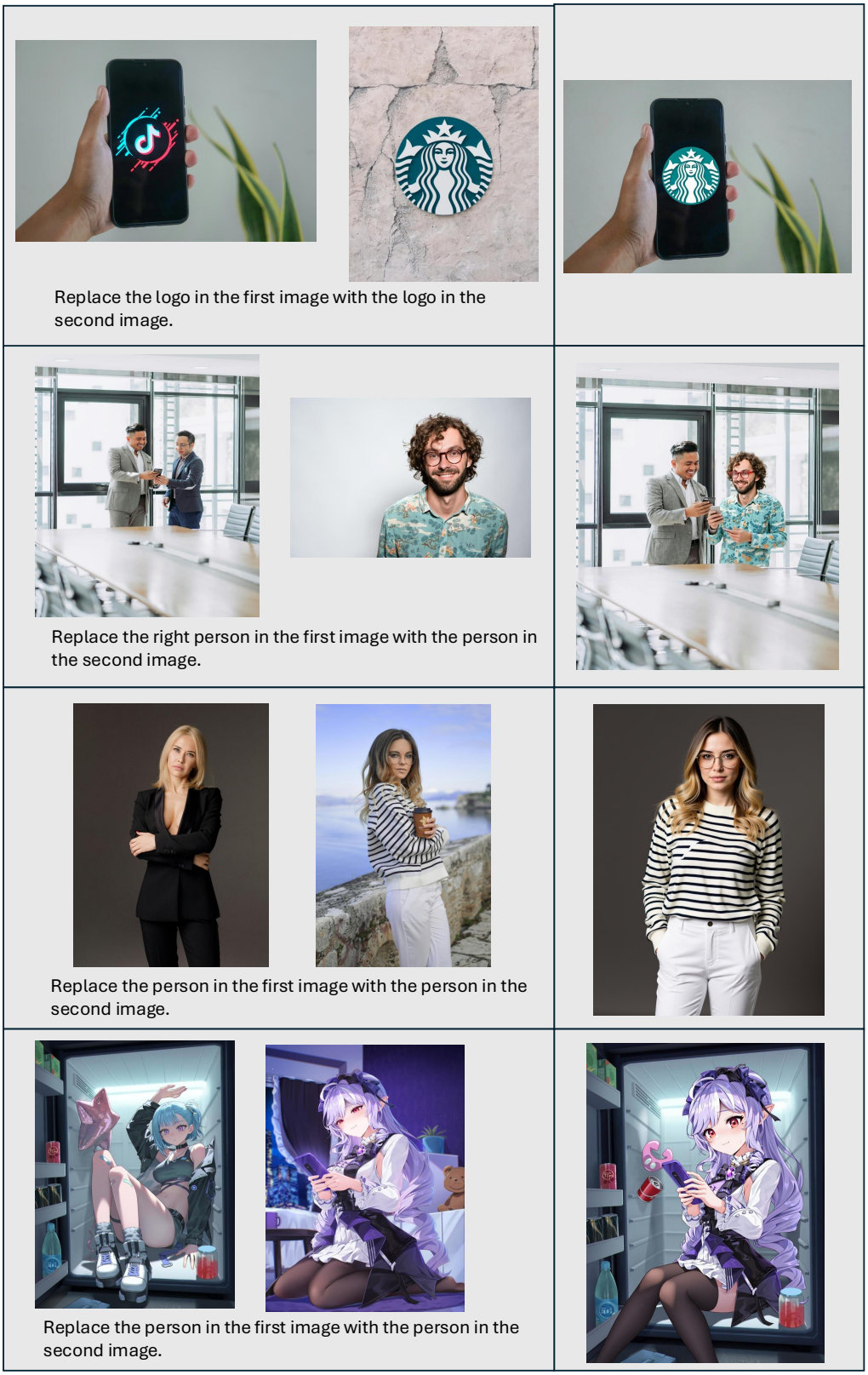


Figure 8. Multimodal instruction-based editing cases of DreamOmni2.

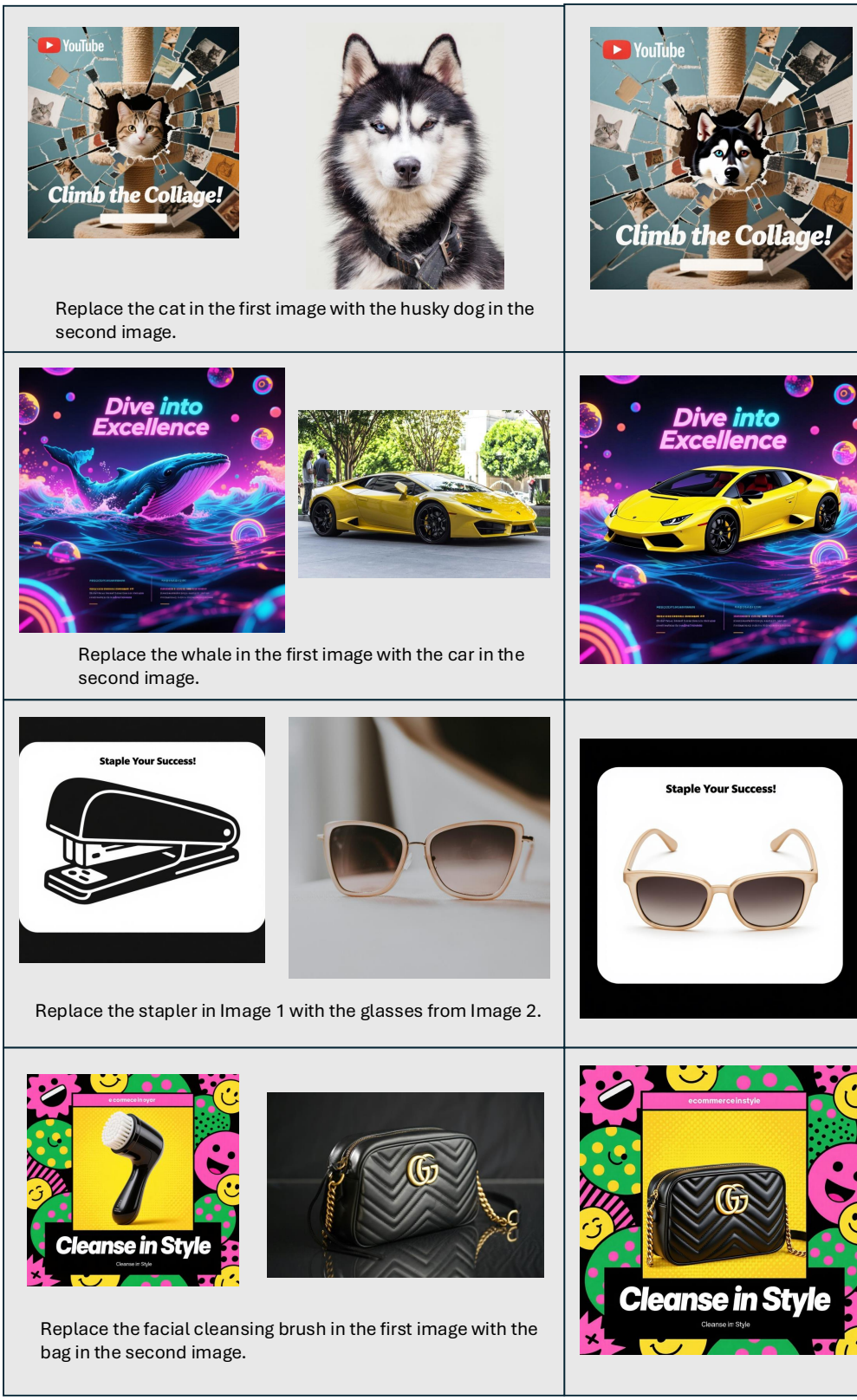


Figure 9. Multimodal instruction-based editing cases of DreamOmni2.

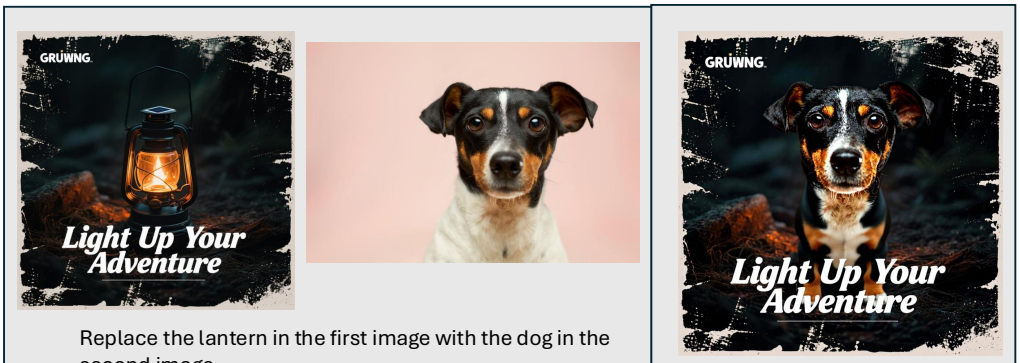


Figure 10. Multimodal instruction-based editing cases of DreamOmni2.

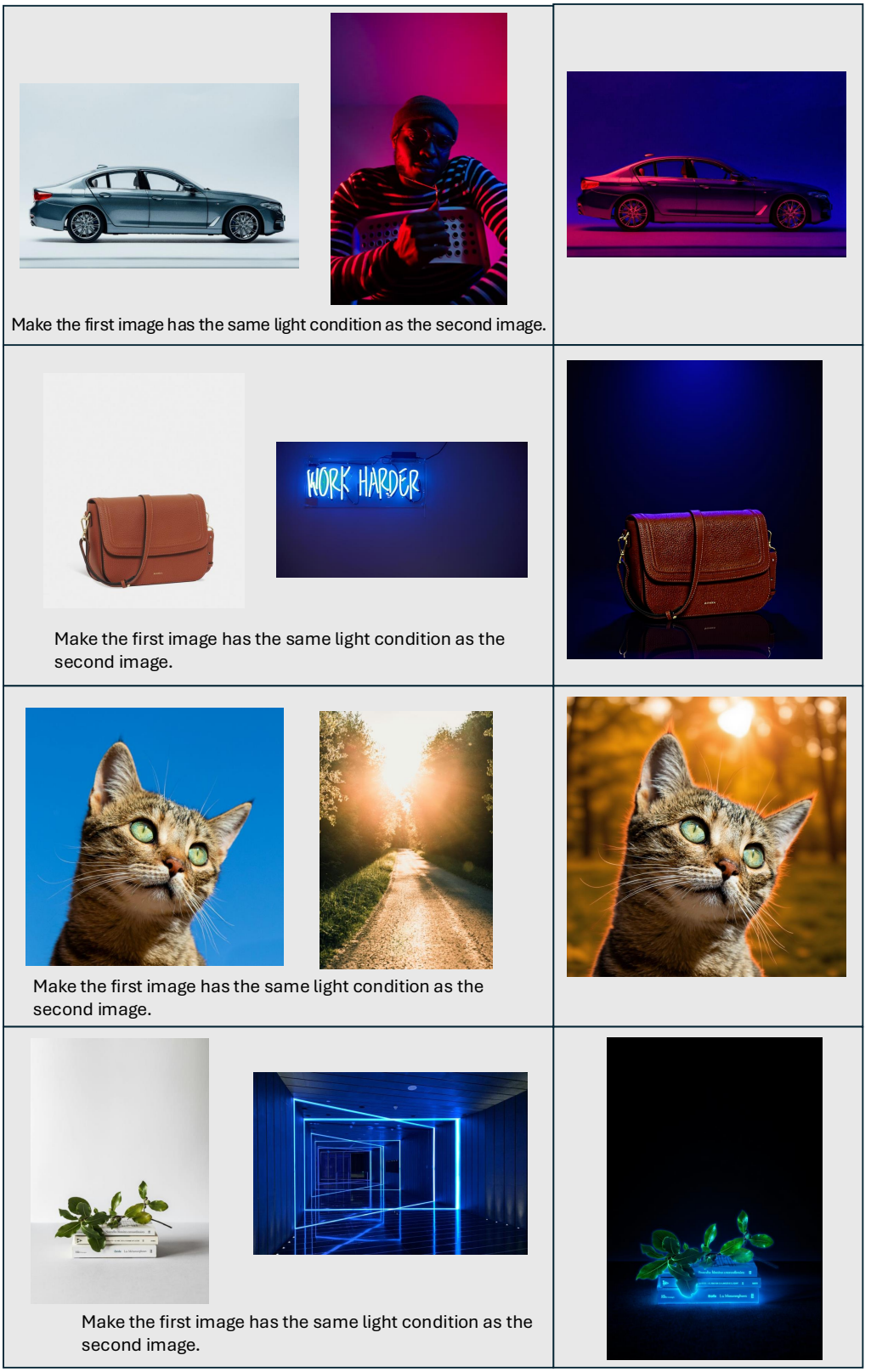


Figure 11. Multimodal instruction-based editing cases of DreamOmni2.

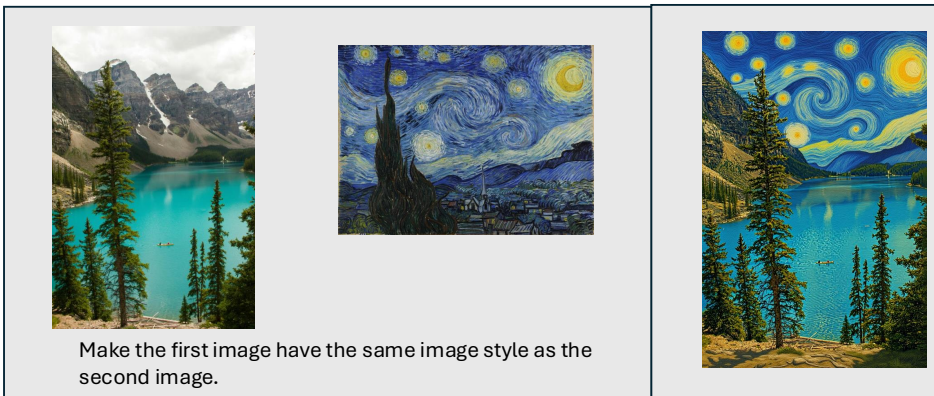


Figure 12. Multimodal instruction-based editing cases of DreamOmni2.

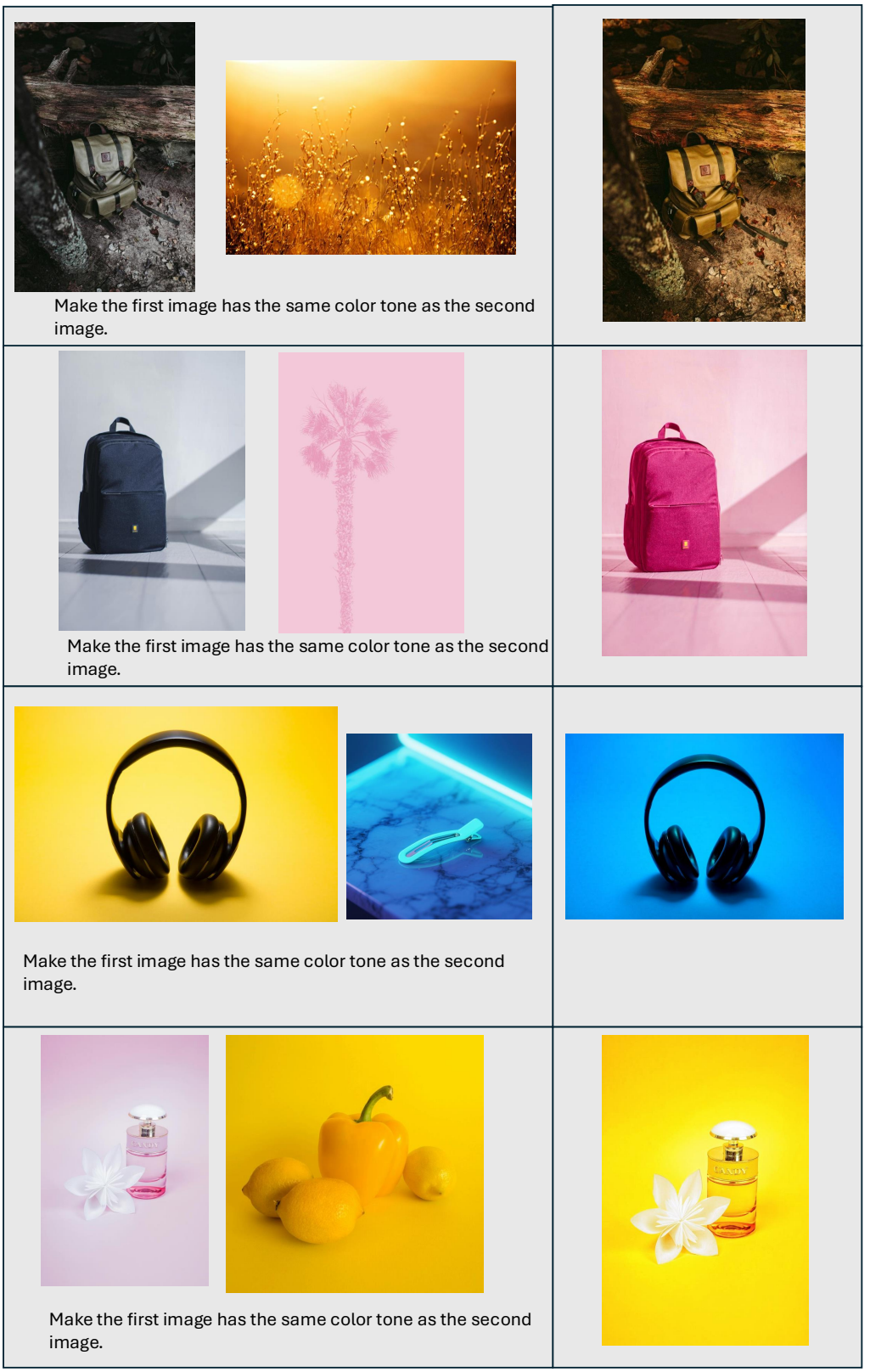


Figure 13. Multimodal instruction-based editing cases of DreamOmni2.

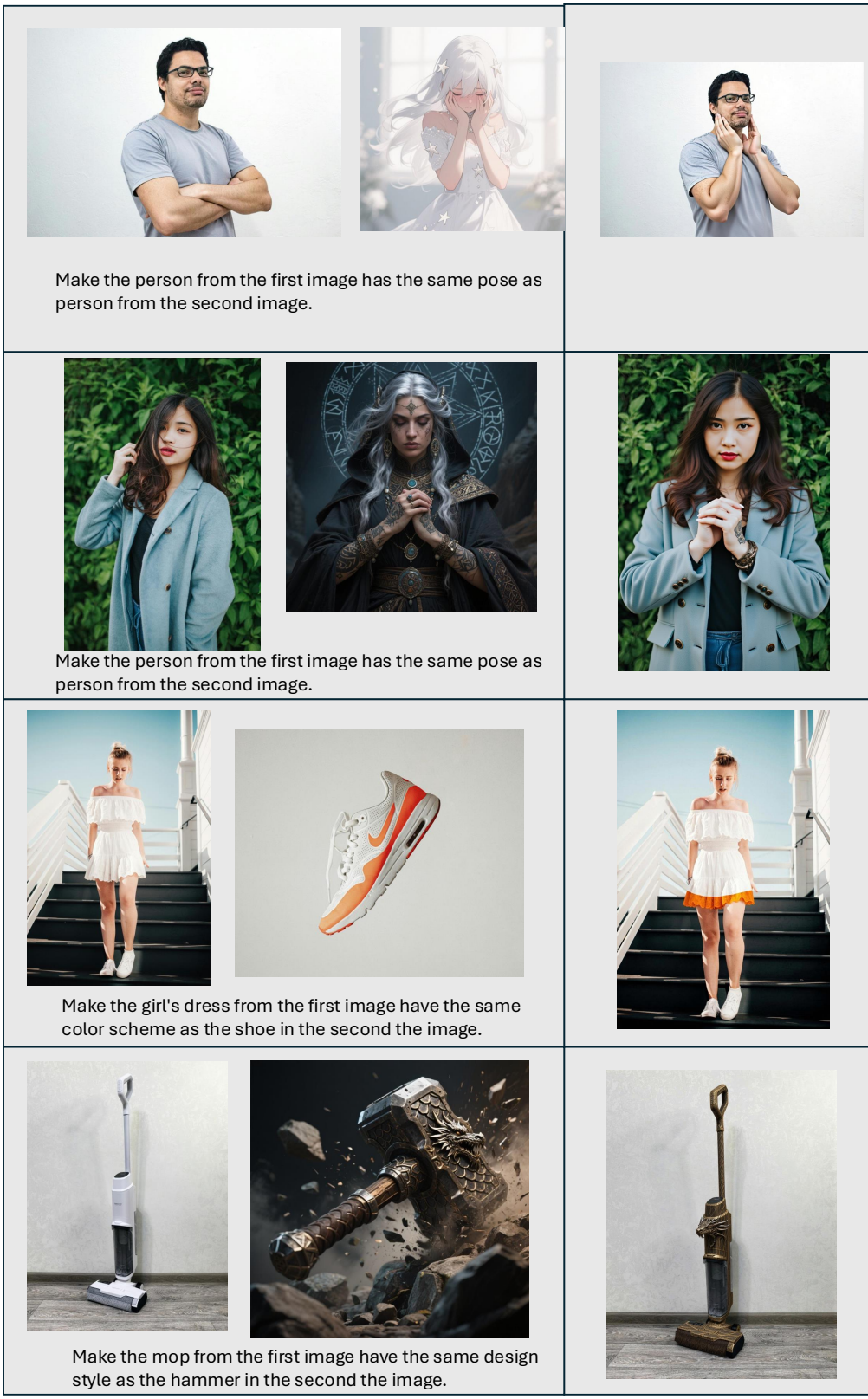


Figure 14. Multimodal instruction-based editing cases of DreamOmni2.

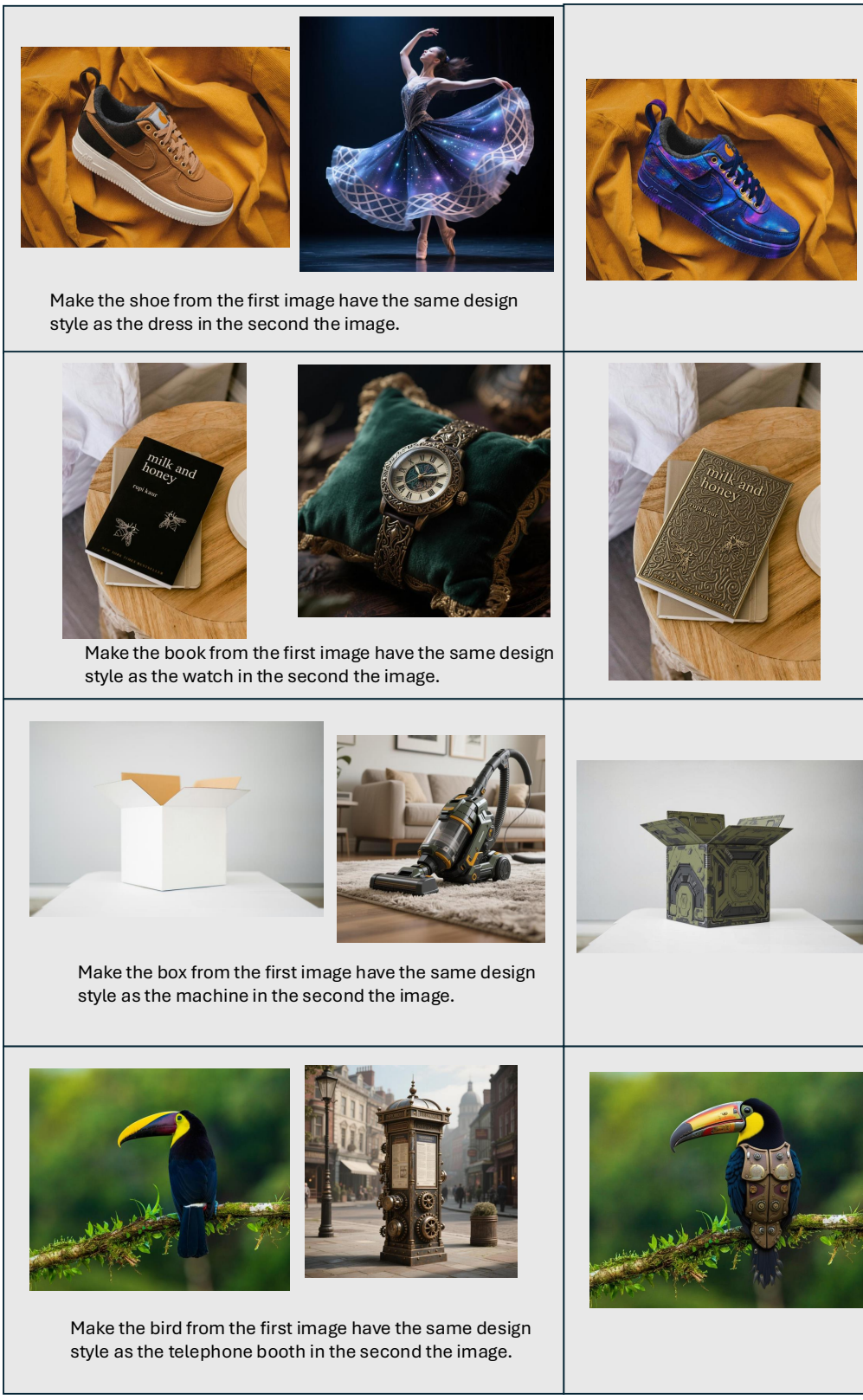


Figure 15. Multimodal instruction-based editing cases of DreamOmni2.

Figure 16. Multimodal instruction-based editing cases of DreamOmni2.

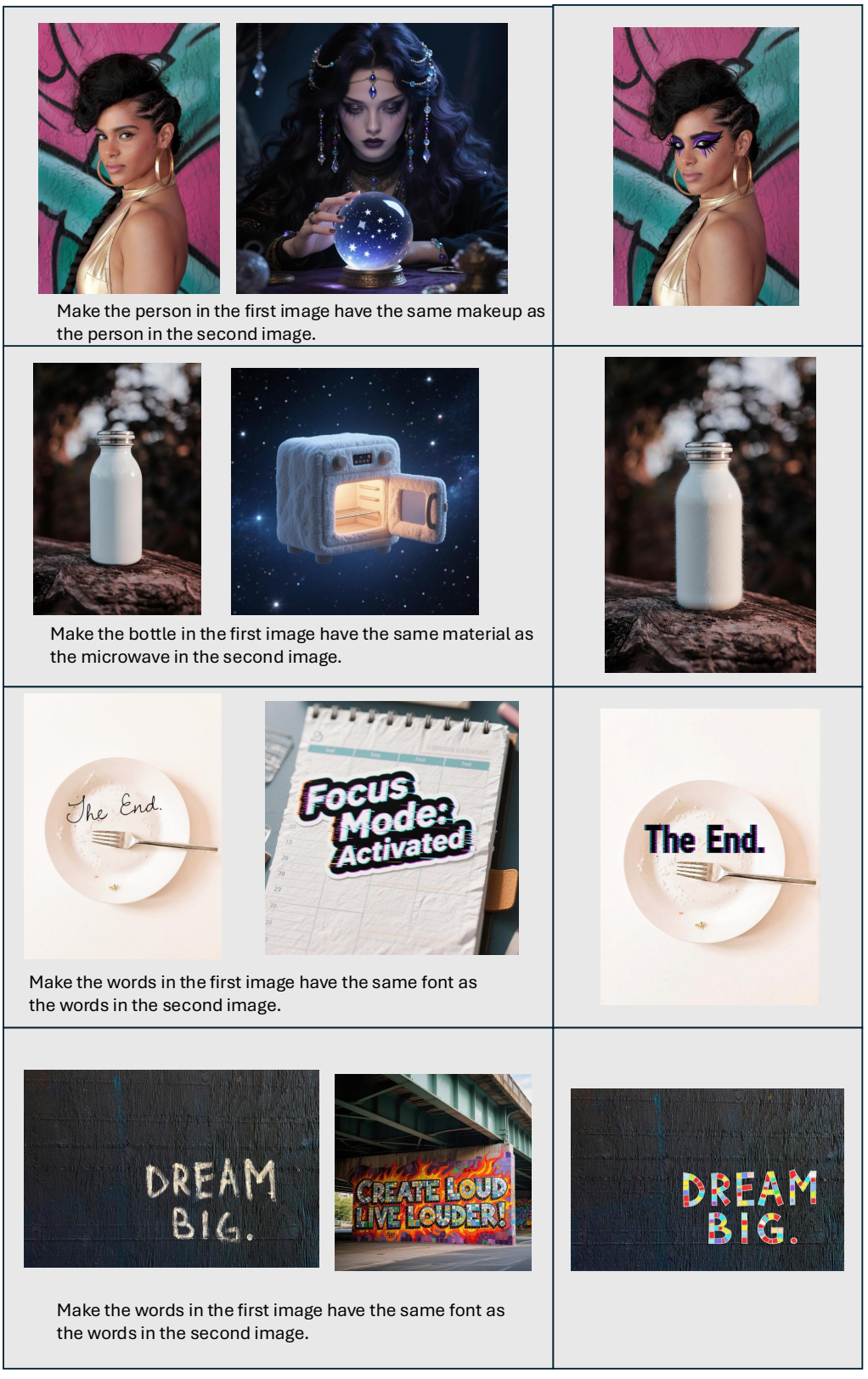
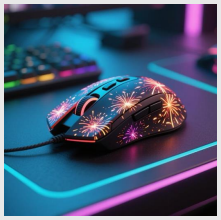
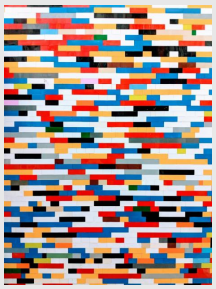
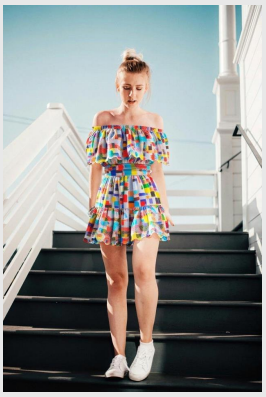


Figure 17. Multimodal instruction-based editing cases of DreamOmni2.


Make the car in the first image have the same pattern as the mouse in the second image.

Make the dress in the first image have the same pattern in the second image.



Make the T-shirt in the first image have the same pattern in the second image.



Make the paper in the first image have the same pattern in the second image.

Figure 18. Multimodal instruction-based editing cases of DreamOmni2.

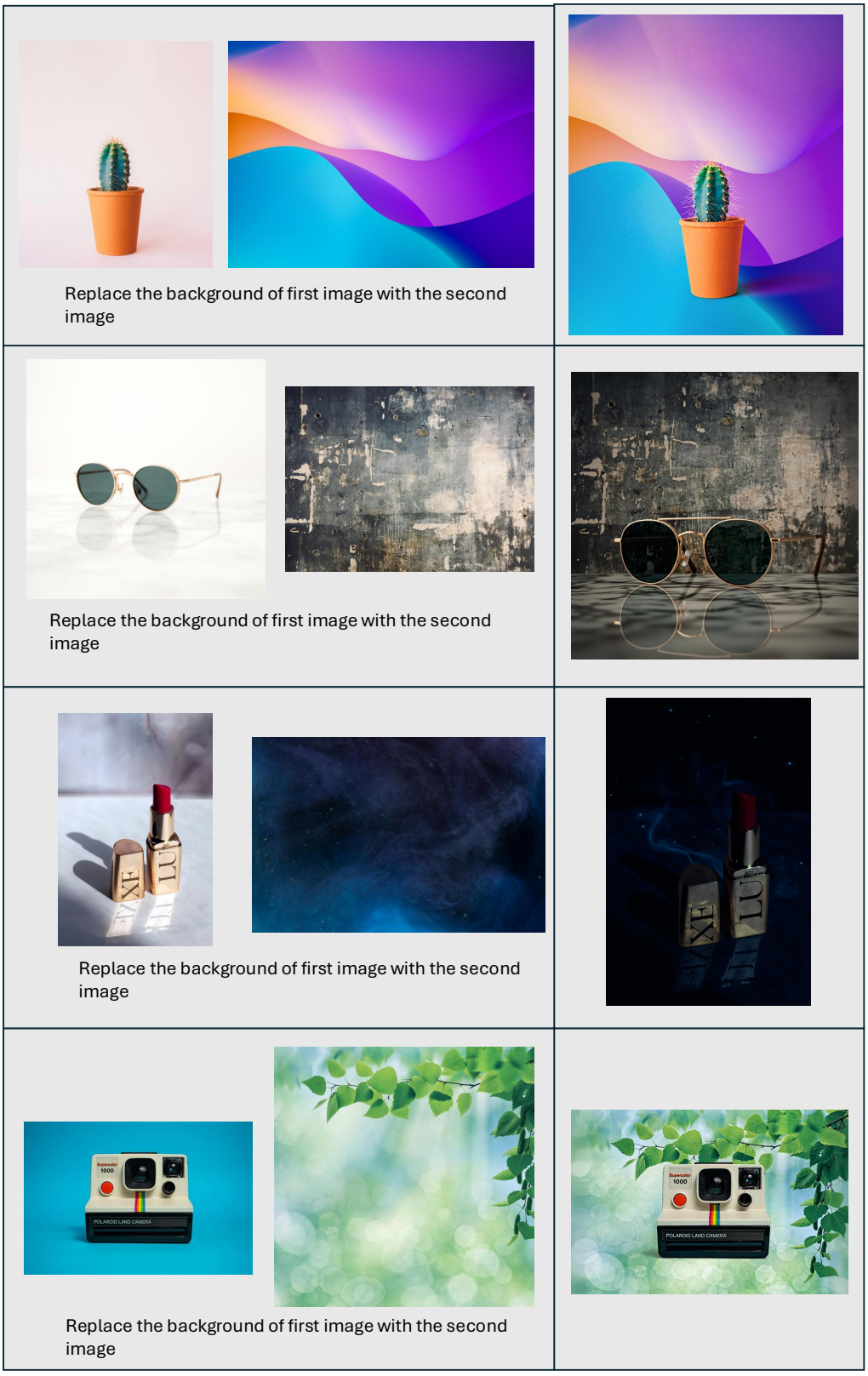
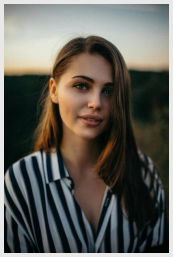
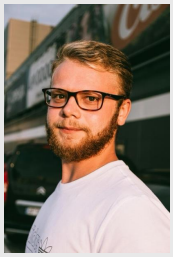
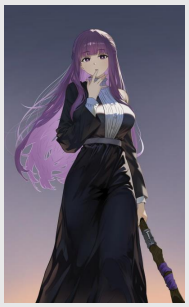


Figure 19. Multimodal instruction-based editing cases of DreamOmni2.

In the scene, the character from the first image stands on the left, and the character from the second image stands on the right. They are shaking hands against the backdrop of a spaceship interior.

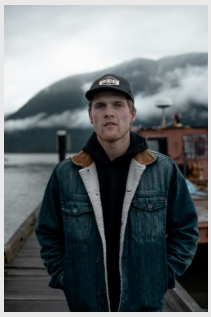



The character from the first image is holding the item from the second picture.

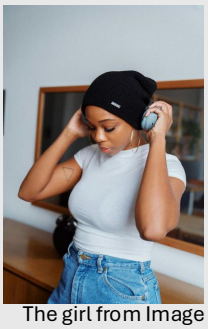
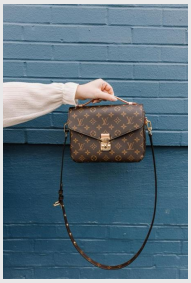

The character from the second image is holding the item from the first image.


On a lawn, the woman from the first image is sitting on the grass, and the backpack from the second image is placed beside her.

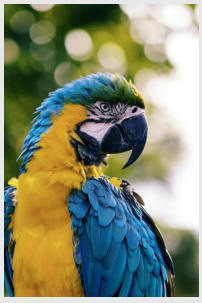
Figure 20. Multimodal instruction-based generation cases of DreamOmni2.




The man from the first image is wearing the clothes from the second image and is sitting on a sofa.

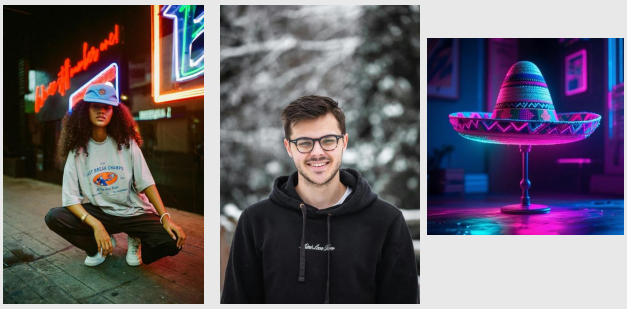

The girl from Image 1 is holding the bag from Image 2 and standing in a bedroom. The color tone of the image is the same as in Image 3.



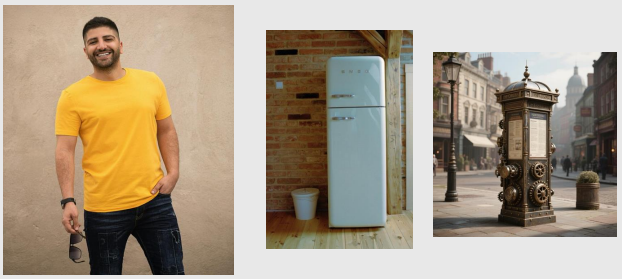
The parrot from Image 1 is wearing the hat from Image 2 and standing on the ground, with a forest in the background. The color tone of the image is the same as in Image 3.


The car from Image 1 is parked on the street, and the bird from Image 2 is standing on the roof of the car. The style of the image is the same as in Image 3.

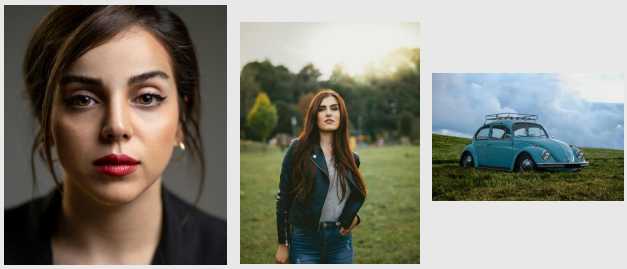
Figure 21. Multimodal instruction-based generation cases of DreamOmni2.



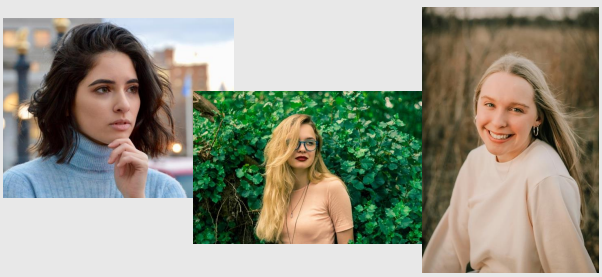
The man from Image 1 and the woman from Image 2 are standing in a laboratory. The light condition of the image is the same as in Image 3.



The man in Figure 1 is sitting on a sofa in a room, next to a refrigerator from Figure 2. The refrigerator's design style matches the object shown in Figure 3.

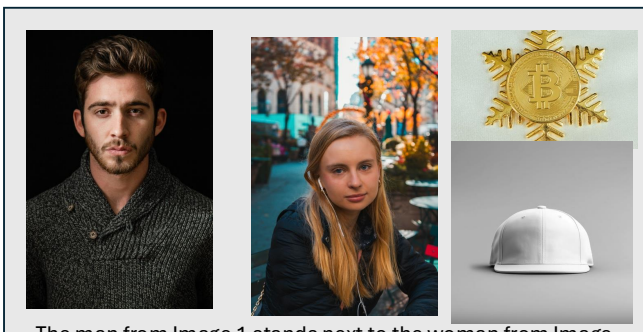


The girl from Image 1 and the girl from Image 2 are standing in front of the car from Image 3.

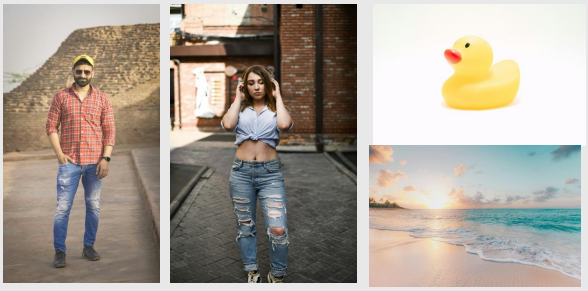


The woman from Image 1 is sitting on the chair, the woman from Image 2 is standing on the left, and the woman from Image 3 is standing on the right. The background is in a living room.

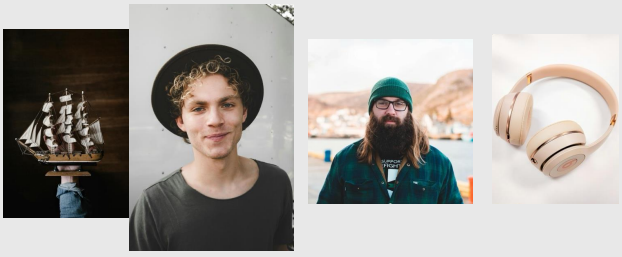
Figure 22. Multimodal instruction-based generation cases of DreamOmni2.



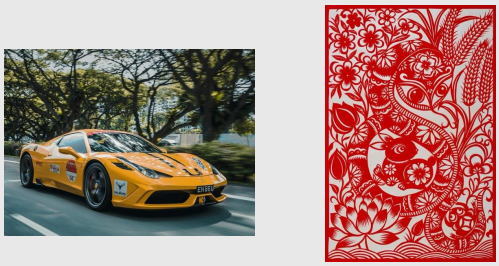
The man from Image 1 stands next to the woman from Image 2. The woman is wearing the hat from Image 4, which has the logo from Image 3 on it. The background is by the lake.



The man in image 1 is standing and holding the toy from image 3. The woman from image 2 is standing next to him. The background is from image 4.



The man from Image 2 is holding the ship model from Image 1, and the man from Image 3 is wearing the headphones from Image 4. They are in a coffee shop.



The car in image 1 is driving through the city, surrounded by tall buildings. The style of the image is the same as in image 2.

Figure 23. Multimodal instruction-based generation cases of DreamOmni2.

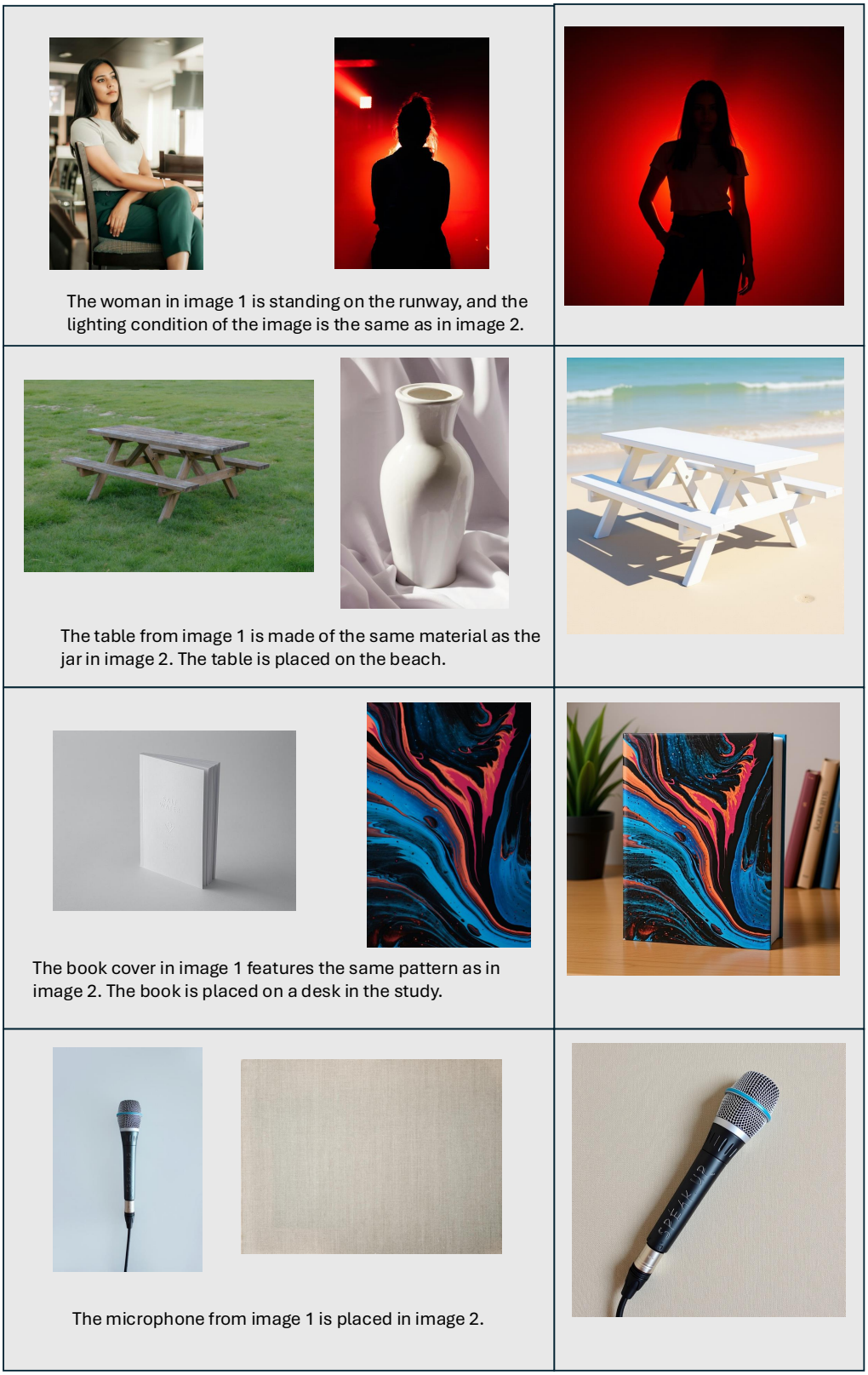
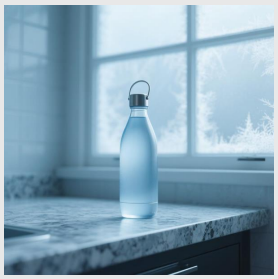
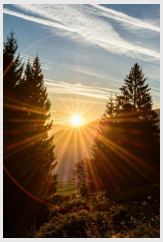


Figure 24. Multimodal instruction-based generation cases of DreamOmni2.

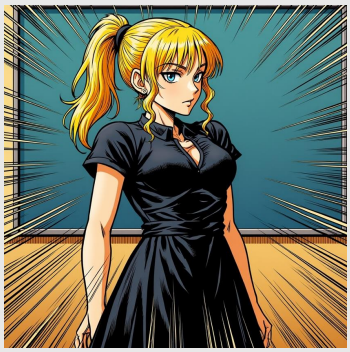


A majestic lion resting on a rocky outcrop. The color tone of the image is the same as in Image 1.



A warrior stands on the battlefield. The lighting conditions of the image are the same as in the reference image.

A blonde girl with a high ponytail, wearing a black long dress, stands at the front of the classroom. The style of the image is the same as the given image.

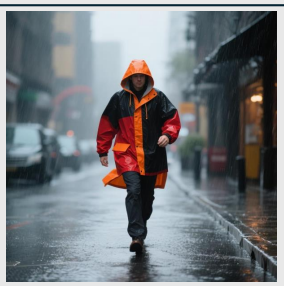


A spaceship is flying in the sky, with the sun visible in the background. The style of the image is the same as in Image 1.

Figure 25. Multimodal instruction-based generation cases of DreamOmni2.

A girl wearing a pink skirt and a white long-sleeve shirt, with long golden hair. She strikes the same pose as the man in the given image. The background is a field of flowers.

Generate a helicopter soaring above a city skyline at dusk. The color scheme of the helicopter is the same as that of the motorcycle.

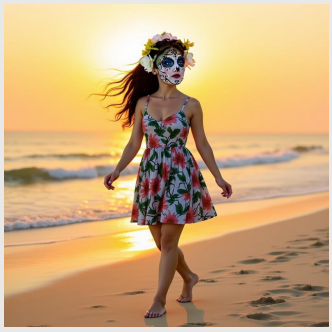


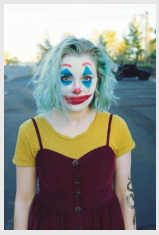
A sleek smartphone resting on a table, with its design featuring smooth curves and a modern look. The color scheme of the phone matches the outfit of the man in the reference image.

A stylish hairdryer placed on a vanity table. The design style of the hairdryer is inspired by the comb in the given image.

Figure 26. Multimodal instruction-based generation cases of DreamOmni2.

Generate a woman in a floral summer dress, walking barefoot along a beach at sunset. Her hair flows in the breeze, and she smiles softly as she watches the waves. Her makeup is the same as the woman in the given image. The background features a golden sun dipping below the horizon, casting warm hues over the calm ocean and sandy shore.





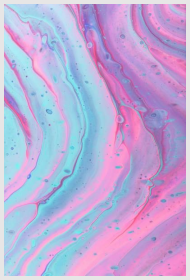
A woman in a cozy knitted sweater and denim jeans, sitting by a fireplace, sipping tea while reading a book. Her hair is styled in loose waves, and she has a calm, content expression. Her makeup is the same as the woman in the given image.

A vintage motorcycle is parked on a cobblestone street. The material of the motorcycle is the same as the vase.

On the cup, "Story" is displayed in the same font style as the reference image.

Figure 27. Multimodal instruction-based generation cases of DreamOmni2.

A table with the texture matching the one in the given image, placed in the center of a cozy living room. The table is surrounded by comfortable chairs, with soft lighting coming from a nearby lamp.



Generate a bottle placed on a wooden kitchen counter. Its texture matches the given image. The bottle is surrounded by fresh fruit, a small bowl of herbs, and a cup of steaming tea.

Figure 28. Multimodal instruction-based generation cases of DreamOmni2.