Pinco: Position-induced Consistent Adapter for Diffusion Transformer in Foreground-conditioned Inpainting (Supplementary Material)

Guangben Lu[1,2*] Yuzhen Du[1,2*] Yizhe Tang[1] Zhimin Sun[1,2] Ran Yi[1†]
Yifan Qi[2] Tianyi Wang[2] Lizhuang Ma[1] Fangyuan Zou[2]
[1]Shanghai Jiao Tong University [2]Tencent

1. Overview

In this supplementary material, we mainly present the following components:

- More implementation details of our model structure and training, and more qualitative results in Sec. 2.
- More ablation study on the convergence of training in Sec. 3.
- More details and cases of the user study in Sec. 4.
- More qualitative comparisons between our Pinco and the state-of-the-art methods in Sec. 5.
- More details and results of the GPT-4o rationality analysis in Sec. 6.
- More inference results under special cases in Sec. 7.
- Limitations in Sec. 8
- Image copyright in Sec. 9

2. Implementation Details

2.1. Model Architecture

Backbone. We apply our proposed Pinco on two DiT-based models, Hunyuan-DiT and Flux-DiT models. For Hunyuan-DiT [5], we use the $DiT-g/2$ config which consists of 40 blocks and has a $1,408$ embed_dim. For Flux-DiT, our Pinco is applied to both the DoubleStreamBlocks and the SingleStreamBlocks. The architecture of Flux-Pinco is shown in Fig. S1.

Decoupled Image Feature Extractor. We use the VAE Encoder of the original DiT model as our semantic feature extractor and only take the output of the final layer as our semantic feature. Meanwhile, we also construct a simple convolutional network (ConvNet) to extract the shape feature. More precisely, since Hunyuan-DiT draws inspiration from the ideas of U-ViT [1] by using skip connections to link the blocks of DiT, we believe that the presence of skip connections allows different blocks to have varying granularities. If we directly feed the output of the last layer of the ConvNet into all modules without distinction, it would weaken the model's inherent perception of feature granularity. Therefore, we extract the outputs from different layers of the ConvNet for different blocks. Specifically:

- The ConvNet consists of 7 convolutional layers, with the outputs of the 1st, 3rd, 5th, and 7th convolutional layers serving as features.
- For Hunyuan-DiT, we divided the 40 blocks into 8 groups, with each group containing 5 blocks. Considering the skip connections, we feed the features from the first layer of the ConvNet into blocks 1-5 and 36-40, and the features from the second layer into blocks 6-10 and 31-35, and so on.
- For Flux-DiT, we apply our pinco on both the 19 DoubleStreamBlocks and the 38 SingleStreamBlocks. And the use of the ConvNet is the same to the HY-Pinco.

Self-Consistent Adapter. We construct corresponding adapters for each block to inject the subject feature. Each adapter first rearranges the feature shape to match the intrinsic latent shape [9] and then uses a linear layer to transform features from a dimension of dim to a dimension of the latent feature. Then, it uses two independent matrices to obtain K and V, which are used to compute the subject-aware attention. The Q is directly taken from the model's computation of self-attention.

2.2. Multi-Aspect Ratio Training

Due to the varying proportions of subjects within the frame (e.g., a vehicle that may occupy more than 50% of the frame, while an item like a shoe might only take up about 10%), it is essential for the model to effectively handle subjects occupying different proportions of the frame and generate a suitable background for subjects of varying sizes. To achieve this, we employ a multi-aspect ratio augmentation method to construct training samples throughout the training process. As illustrated in Fig. S2, for each high-quality image, we define the minimum bounding rectangle of the subject based on the mask (in red), while the bounding rectangle of the entire image serves as the maximum range (in blue). We designate the areas of the maximum and minimum frames as the upper and lower bounds of a normal

[*] Equal contribution.
[†] Corresponding author. ranyi@sjtu.edu.cn

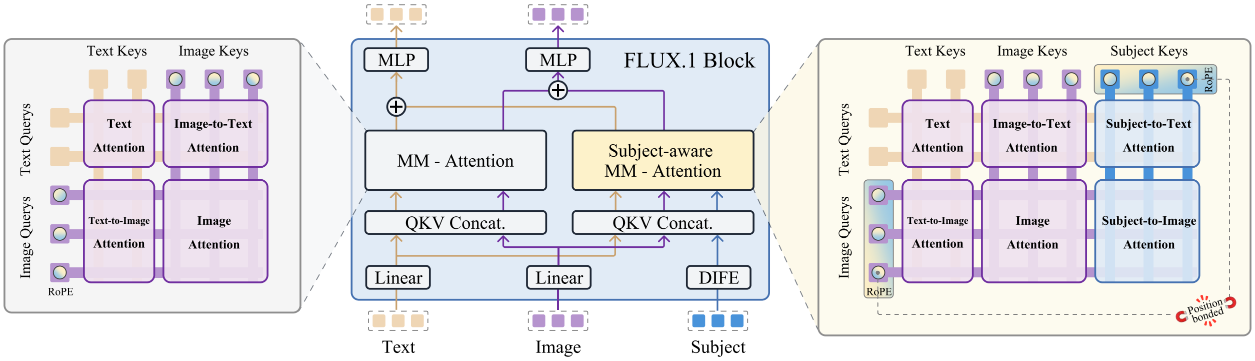


Figure S1. The architecture of Flux-Pinco. For MM-DiT, we concatenate the latent and the subject features together to calculate the subject-aware attention.

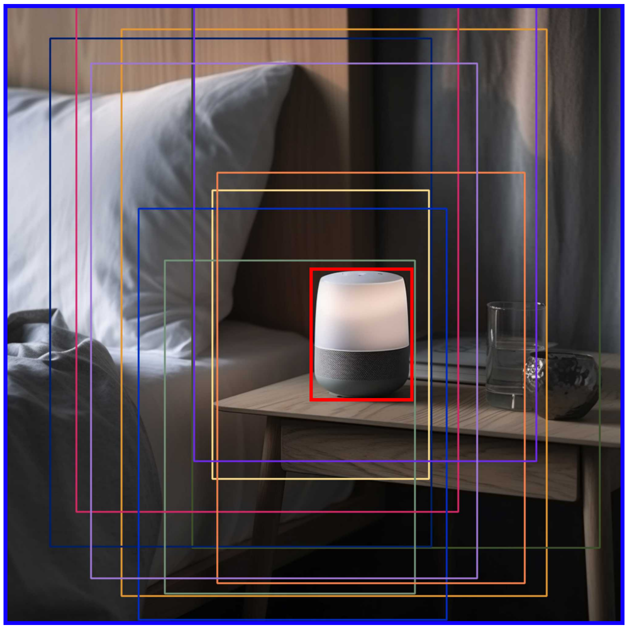


Figure S2. Method for obtaining multi-aspect ratio samples.

distribution, respectively. During training, we sample various shapes and locations of bounding rectangles to create training samples with diverse frame proportions and aspect ratios (e.g., 16:9, 9:16, 1:1). Fig. S6 shows more cases generated by Pinco with different aspect ratios.

3. More Ablation Study

3.1. Convergence Process Analysis

Fig. S4 shows the convergence analysis of HY-Pinco, Pinco-w/oRoPE, and Pinco-Cross. The two images illustrate the changes in image OER and DINO similarity within the mask area as the training epochs extended. Evidently, without the aid of shared positional embedding anchor, the model struggles to effectively incorporate subject features, resulting in consistently poor scores for both OER and DINO similarity. This highlights the importance of shared positional encoding in effectively utilizing subject features and ensuring the consistency of the subjects in the generated images. On the other hand, injecting subject features in the self-attention layer leads to faster convergence during training, with lower OER and better DINO similarity compared to injecting features in the cross-attention layer. This further supports the rationale and effectiveness of injecting features in the self-attention layer. Fig. S7 demonstrates more cases between Pinco and Pinco-w/oRoPE during the training process.

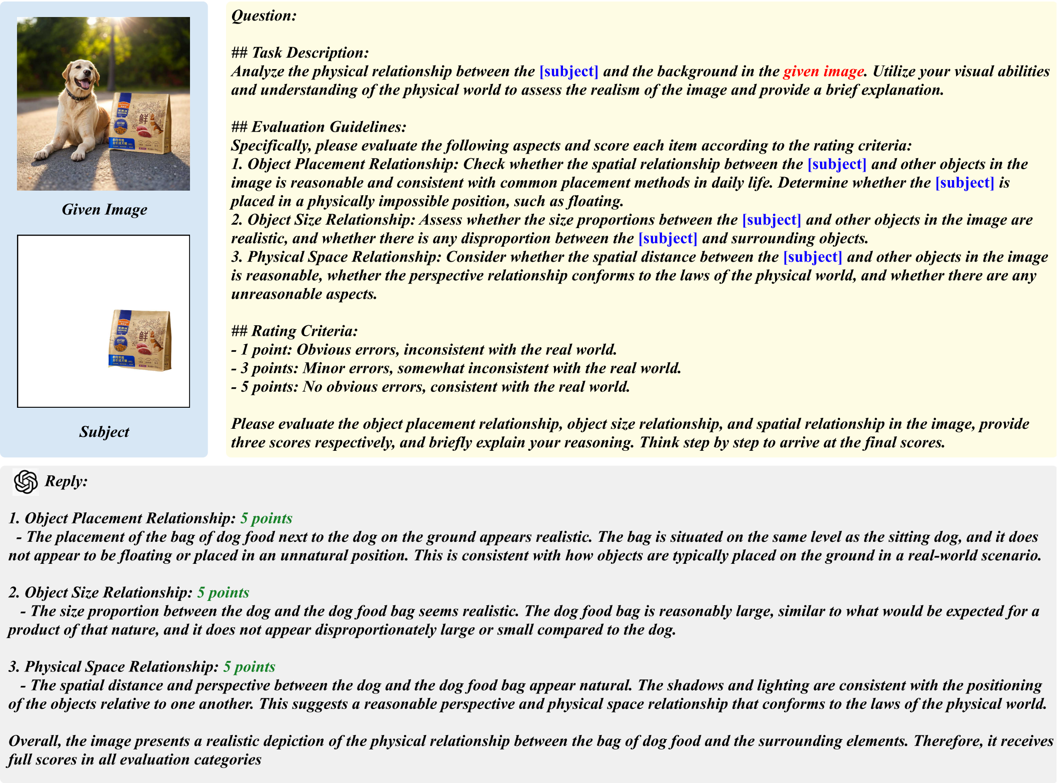
4. User Study

During the user study, participants were asked to evaluate side-by-side samples from multiple aspects, including the rationality of the background, the appropriateness of object sizes, the suitability of object placements, and the harmony between the subjects and the background, and select the images they considered to be better. Fig. S8 displays some cases from the user study.

5. More Qualitative Comparisons

We provide more qualitative comparisons between our Pinco and the state-of-the-art methods in Fig. S9, S10 and S11. The compared baselines include:

- SD1.5 backbone: ControlNet inpainting [10], HD-Painter [6], PowerPaint [11], and BrushNet-SD1.5 [3];
- SDXL backbone: SDXL inpainting [7], layerdiffusion [4], BrushNet-SDXL [3], and Kolors-inpainting [8];
- DiT-based models: HY-ControlNet (Hunyuan-DiT backbone), and Flux ControlNet [2] (FLUX.1 backbone).



Given Image

Subject

Question:

## Task Description:
Analyze the physical relationship between the [subject] and the background in the given image. Utilize your visual abilities and understanding of the physical world to assess the realism of the image and provide a brief explanation.

## Evaluation Guidelines:
Specifically, please evaluate the following aspects and score each item according to the rating criteria:
1. Object Placement Relationship: Check whether the spatial relationship between the [subject] and other objects in the image is reasonable and consistent with common placement methods in daily life. Determine whether the [subject] is placed in a physically impossible position, such as floating.
2. Object Size Relationship: Assess whether the size proportions between the [subject] and other objects in the image are realistic, and whether there is any disproportion between the [subject] and surrounding objects.
3. Physical Space Relationship: Consider whether the spatial distance between the [subject] and other objects in the image is reasonable, whether the perspective relationship conforms to the laws of the physical world, and whether there are any unreasonable aspects.

## Rating Criteria:
- 1 point: Obvious errors, inconsistent with the real world.
- 3 points: Minor errors, somewhat inconsistent with the real world.
- 5 points: No obvious errors, consistent with the real world.

Please evaluate the object placement relationship, object size relationship, and spatial relationship in the image, provide three scores respectively, and briefly explain your reasoning. Think step by step to arrive at the final scores.

Reply:

1. Object Placement Relationship: 5 points
- The placement of the bag of dog food next to the dog on the ground appears realistic. The bag is situated on the same level as the sitting dog, and it does not appear to be floating or placed in an unnatural position. This is consistent with how objects are typically placed on the ground in a real-world scenario.

2. Object Size Relationship: 5 points
- The size proportion between the dog and the dog food bag seems realistic. The dog food bag is reasonably large, similar to what would be expected for a product of that nature, and it does not appear disproportionately large or small compared to the dog.

3. Physical Space Relationship: 5 points
- The spatial distance and perspective between the dog and the dog food bag appear natural. The shadows and lighting are consistent with the positioning of the objects relative to one another. This suggests a reasonable perspective and physical space relationship that conforms to the laws of the physical world.

Overall, the image presents a realistic depiction of the physical relationship between the bag of dog food and the surrounding elements. Therefore, it receives full scores in all evaluation categories

Figure S3. GPT-4o prompt for assessment and its reply. Note that you need to specify the name of the subject in [subject]

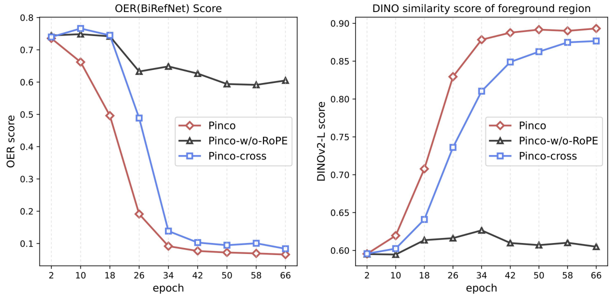


Figure S4. The convergence analysis of Pinco, Pinco-Cross, and Pinco-w/oRoPE. Pinco-Self can maintain better shape constraints and foreground consistency while achieving efficient training.

6. GPT-4o Rationality

We leverage the GPT-4o to evaluate each image based on Object Placement Relationship, Object Size Relationship, and Physical Space Relationship. The criteria for these three aspects and rating criteria are as follows:

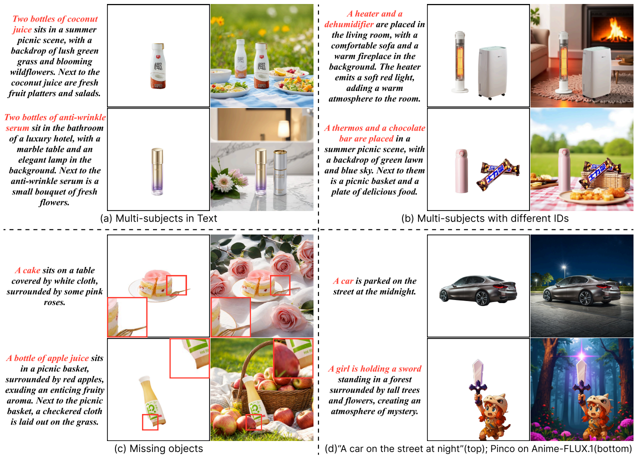
- Object Placement Relationship: Check whether the spatial relationship between the subject and other objects in the image is reasonable and consistent with common placement methods in daily life. Determine if the subject is placed in a physically impossible position, such as floating.
- Object Size Relationship: Assess whether the size proportions between the subject and other objects in the image are realistic and whether there is any disproportion between the subject and surrounding objects.
- Physical Space Relationship: Consider whether the spatial distance between the subject and other objects in the image is reasonable, whether the perspective relationship conforms to the laws of the physical world, and whether there are any unreasonable aspects.
- Rating Criteria: 1 point: Obvious errors, inconsistent with the real world. 3 points: Minor errors, somewhat inconsistent with the real world. 5 points: No obvious errors, consistent with the real world.

Fig. S12 presents the results of the rationality analysis for the images returned by GPT-4o under the criteria mentioned above. The detailed prompt given to GPT-4o and the reply are shown in Fig. S3.

7. Inference Results under Special Cases.

To verify the robustness of our method in the inference phase, we conducted the following special experiments:

Text-image Interdependence. We test the case where the number of subjects in the text description is greater than the number of subjects in the conditional image. As shown in Fig. S5 (a), our Pinco will generate another subject which is



Two bottles of coconut juice sits in a summer picnic scene, with a backdrop of lush green grass and blooming wildflowers. Next to the coconut juice are fresh fruit platters and salads.

Two bottles of anti-wrinkle serum sit in the bathroom of a luxury hotel, with a marble table and an elegant lamp in the background. Next to the anti-wrinkle serum is a small bouquet of fresh flowers.

(a) Multi-subjects in Text

A heater and a dehumidifier are placed in the living room, with a comfortable sofa and a warm fireplace in the background. The heater emits a soft red light, adding a warm atmosphere to the room.

A thermos and a chocolate bar are placed in a summer picnic scene, with a backdrop of green lawn and blue sky. Next to them is a picnic basket and a plate of delicious food.

(b) Multi-subjects with different IDs

A cake sits on a table covered by white cloth, surrounded by some pink roses.

A bottle of apple juice sits in a picnic basket, surrounded by red apples, exuding an enticing fruity aroma. Next to the picnic basket, a checkered cloth is laid out on the grass.

(c) Missing objects

A car is parked on the street at the midnight.

A girl is holding a sword standing in a forest surrounded by tall trees and flowers, creating an atmosphere of mystery.

(d)"A car on the street at night"(top); Pinco on Anime-FLUX.1(bottom)

Figure S5. Some inference results under special cases.

aligned with the given prompts. It is worth noting that additional subjects appear only when the text describes multiple subjects.

Multiple Subjects with Different IDs. We test the case where there are multiple subjects with different IDs in the conditional foreground image. As shown in Fig. S5 (b), our Pinco can generate perfect results aligned with the given prompts.

Completing Missing Objects. We test the case where the conditional foreground objects have missing parts. As shown in Fig. S5 (c), for partially missing objects, our Pinco either completes the missing foreground parts or uses background objects to cover them for a harmonious result.

Textual Conflicts. We test the case where the conditional foreground image and the textual background description conflict in lighting conditions. In such case, the model will adjust the background to better fit the foreground while ensuring alignment with the text. For example, given a foreground car in bright scene and the textual description of night, the result might show a car lit by a street lamp on a nighttime street, As shown in Fig. S5 (d) (top).

Plug and Play Property. As an adapter, our Pinco can be applied to different models with the same structure with no need of extra training. For example, we apply our Pinco trained on FLUX.1-dev onto an Anime-style-finetuned FLUX by community. As shown in Fig. S5 (d) (bottom), it can still generate excellent foreground-conditioned inpainting results.

8. Limitations

In foreground-conditioned background inpainting task, one of the most significant requirements is to ensure the foreground consistency. Existing methods generally maintain the internal detailed features of the input foreground well. However, due to their unstable feature injection, the model usually generates some extended parts based on its hallucinations. In Pinco, although we proposed dedicated Self-Consistent Adapter to facilitate a harmonious interaction between foreground features and the overall image layout, sometimes it is still hard to control the shape when dealing with very slender objects like ropes or sticks. In addition, when the input foreground object is captured from an unusual viewpoint, the model may not understand the perspective of the object, resulting in an unreasonable positional relationship between the generated background and foreground.

9. Copyright

Some of the images presented in this paper are sourced from publicly available online resources. In our usage context, we have uniformly retained only the main subject of the original images and removed the background parts. In this section, we specify the exact sources of the images in the form of image links and author credits in Tab. 1. Except for the images explicitly credited with the author and source, all other images are derived from client cases or generated by open-sourced models. The copyright of the images belongs to the original authors and brands. **The images used in this paper are solely for academic research purposes and are only used to test the effectiveness of algorithms. They are not intended for any commercial use or unauthorized distribution.**

For each image containing multiple sub-images, we number them in order from left to right and from top to bottom. For example, in Figure. S6, the sub-images in the first row are numbered as SubFig.S6-1, SubFig.S6-2, and so on, while the sub-images in the second row are numbered as SubFig.S6-5, and so forth.

References

[1] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models, 2023. 1

[2] AliMama Creative. Flux.1-dev controlnet inpainting beta, 2024. Accessed: 2023-10. 2

[3] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. arXiv preprint arXiv:2403.06976, 2024. 2

[4] Pengzhi Li, Qinxuan Huang, Yikang Ding, and Zhiheng Li. Layerdiffusion: Layered controlled image editing with diffusion models. In SIGGRAPH Asia 2023 Technical Communications, pages 1–4. 2023. 2

[5] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748, 2024. 1

Table 1. Copyright of the images in our paper.

Figure	Source
Figure. 1	SubFig.1-2 (ANTA), SubFig.1-3 (METHOU), SubFig.1-5 (BALMUDA), SubFig.1-6 (FLYCO), SubFig.1-7 (Alpha Coders)
Figure. 2	SubFig.2-1 (ZCOOL)
Figure. 3	SubFig.3-1 (Mazda)
Figure. 4	SubFig.4-1 (Foodography), SubFig.4-2 (Xiangma), SubFig.4-3 (IMAXER), SubFig.4-4 (Foodography), SubFig.4-5 (IM Motors), SubFig.4-6 (Foodography)
Figure. 6	SubFig.6-1 (NOW VISION), SubFig.6-2 (Foodography)
Figure. 7	SubFig.7-2/4 (MANTO)
Figure. 8	SubFig.8-1/2 (Foodography), SubFig.8-3/4 (LOTTO)
Figure. S3	SubFig.S3-1/2 (DESING)
Figure. S5	SubFig.S5-a1 (Foodography), SubFig.S5-a2 (Foodography), SubFig.S5-b1/1 (Foodography), SubFig.S5-b1/2 (Foodography), SubFig.S5-b2/1 (Lifease), SubFig.S5-b2/2 (Snickers), SubFig.S5-c1 (Cake), SubFig.S5-c2 (Box Studio), SubFig.S5-d1 (Wallpaper Flare), SubFig.S5-d2 (Feiyu),
Figure. S6	SubFig.S6-1 (MAOOXD), SubFig.S6-2 (Foodography), SubFig.S6-3 (Foodography), SubFig.S6-4 (SUPOR), SubFig.S6-5 (JianmuPhotography), SubFig.S6-7 (Alpha Coders), SubFig.S6-8 (BYHEALTH), SubFig.S6-9 (Rarakiddo), SubFig.S6-10 (Foodography), SubFig.S6-11 (BALMUDA), SubFig.S6-12 (LIBY), SubFig.S6-13 (ROLEX), SubFig.S6-14 (LUXEED), SubFig.S6-15 (L'Oreal), SubFig.S6-16 (Olena Bohovyk), SubFig.S6-17 (BALMUDA), SubFig.S6-18 (BALMUDA)
Figure. S7	SubFig.S7-1 (Xiangma), SubFig.S7-2 (METHOU), SubFig.S7-3 (Foodography), SubFig.S7-4 (NOW VISION), SubFig.S7-5 (Mazda), SubFig.S7-6 (JianmuPhotography), SubFig.S7-7 (ANTA), SubFig.S7-8 (Apple), SubFig.S7-9 (Helen Keller)
Figure. S8	SubFig.S8-1 (FIND VISUAL), SubFig.S8-2 (Nooie Robot Vacuum), SubFig.S8-3 (RIMOWA), SubFig.S8-4 (XIAOMI), SubFig.S8-5 (Box Studio), SubFig.S8-6, SubFig.S8-7 (Foodography), SubFig.S8-8 (FIND VISUAL)
Figure. S9	SubFig.S9-1 (Xiangma), SubFig.S9-2 (METHOU), SubFig.S9-3 (FILMSAYS), SubFig.S9-4 (Alpha Coders)
Figure. S10	SubFig.S10-2 (AUPRES), SubFig.S10-3, SubFig.S10-4 (MAOOXD)
Figure. S11	SubFig.S11-1 (Bear), SubFig.S11-2 (Luckin Coffee), SubFig.S11-3 (Apple), SubFig.S11-4 (FIND VISUAL)
Figure. S12	SubFig.S12-1 (Bear),SubFig.S12-2 (Foodography), SubFig.S12-3 (YUANYI).

[6] Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Hd-painter: high-resolution and prompt-faithful text-guided image inpainting with diffusion models. arXiv preprint arXiv:2312.14091, 2023. 2

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023. 2

[8] Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. arXiv preprint, 2024. 2

[9] Xiangtian Xue, Jiasong Wu, Youyong Kong, Lotfi Senhadji, and Huazhong Shu. St-ldm: A universal framework for text-grounded object generation in real images, 2024. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 2

[11] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. arXiv preprint arXiv:2312.03594, 2023. 2

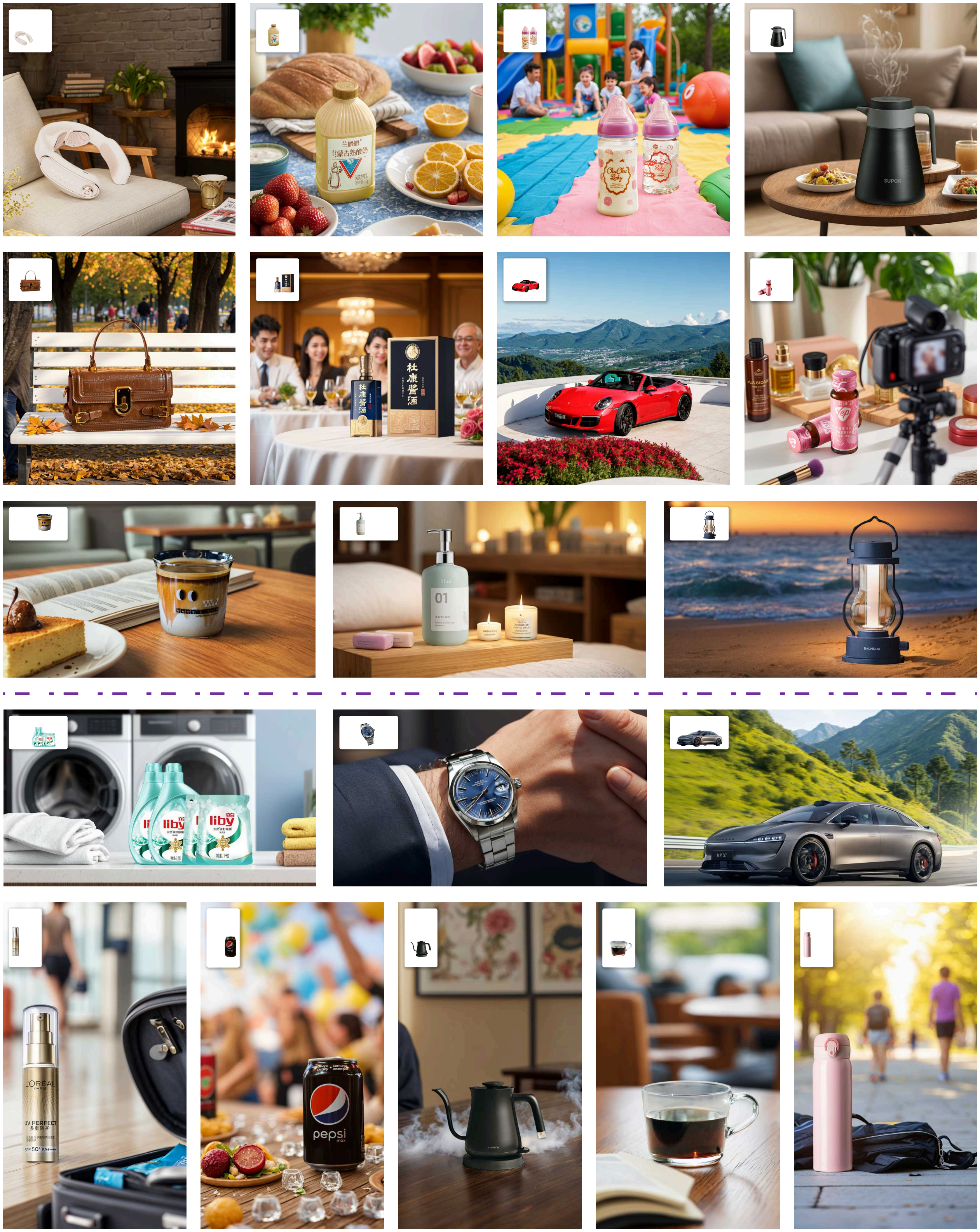


Figure S6. More cases generated by Pinco. Pinco supports the generation of high-quality images with different aspect ratios, while ensuring the reasonable placement of subjects, achieving realistic foreground-conditioned inpainting. The 11 images in the upper section are generated by HY-Pinco, while the 8 images in the lower section are produced by Flux-Pinco.

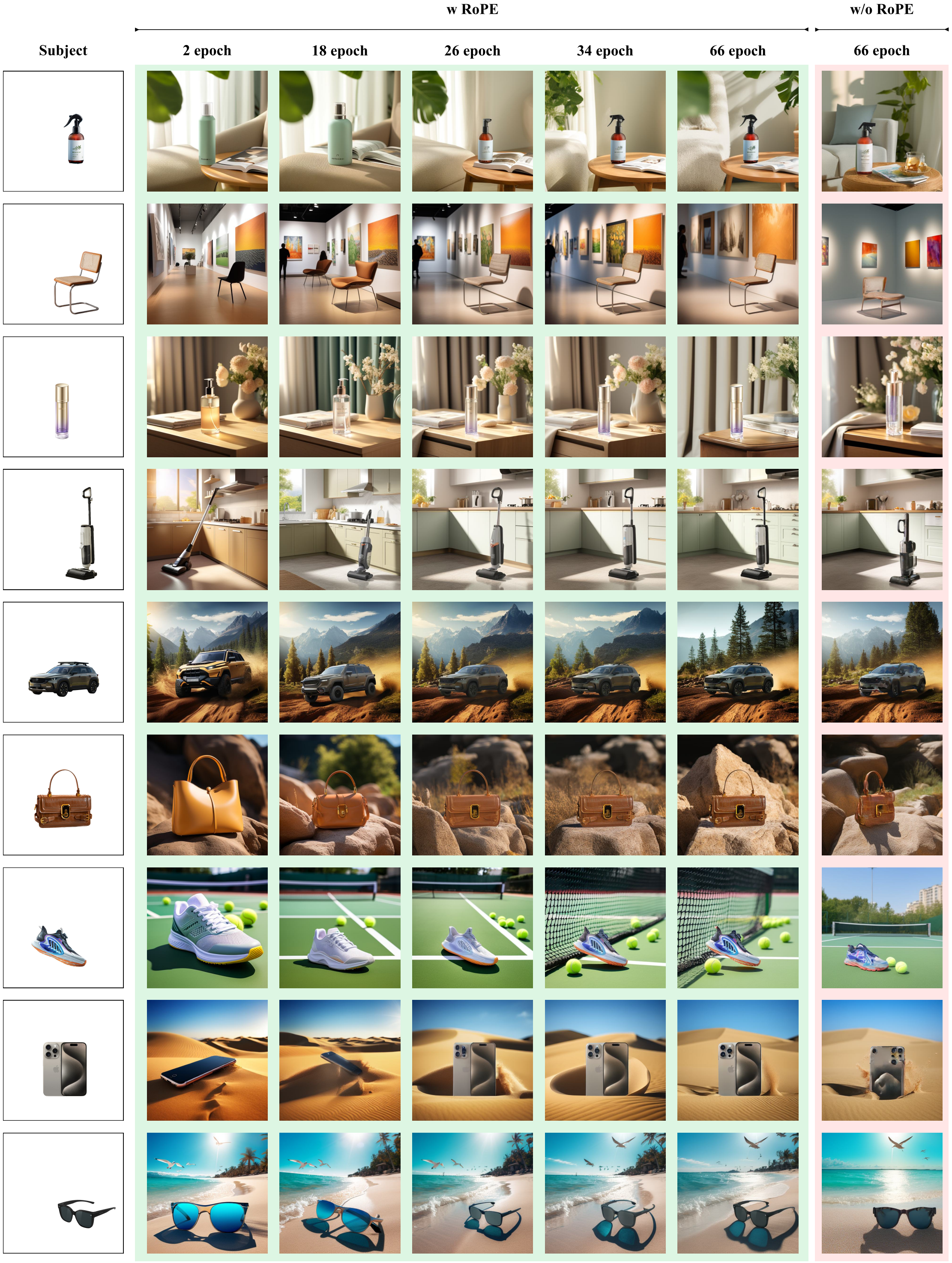
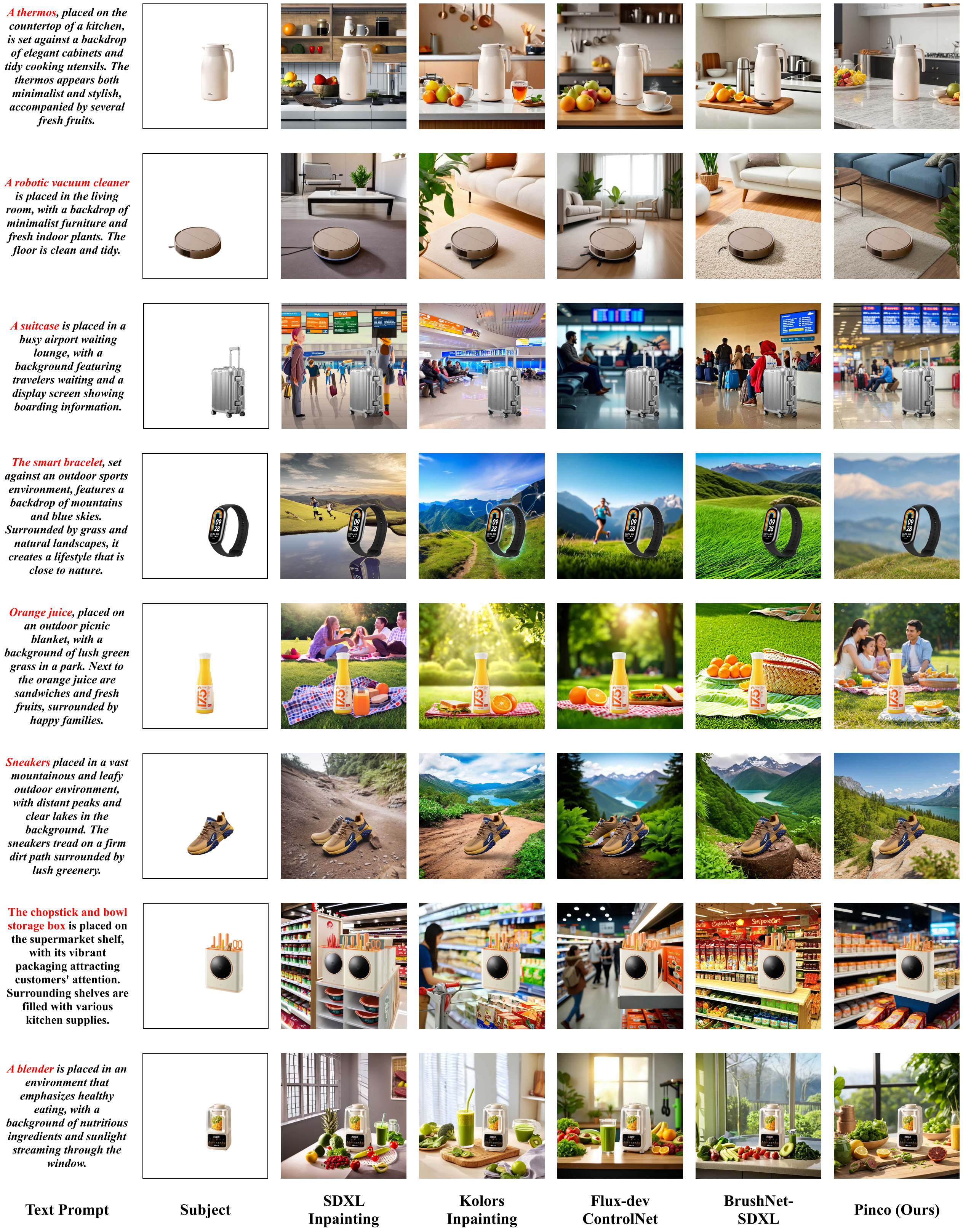


Figure S7. More cases of the model training process. We provide a detailed demonstration of the training process of the Pinco, while only presenting the final results of the Pinco-w/o RoPE due to limited space. We can observe that the Pinco model can gradually place the given subject into the generated scene while maintaining the contour and foreground consistency. In contrast, although the same number of training epochs were used, the Pinco-w/o RoPE model can only learn partial subject information, resulting in distortions in both contour and foreground consistency. Zoom in to observe the details.


A thermos, placed on the countertop of a kitchen, is set against a backdrop of elegant cabinets and tidy cooking utensils. The thermos appears both minimalist and stylish, accompanied by several fresh fruits.

A robotic vacuum cleaner is placed in the living room, with a backdrop of minimalist furniture and fresh indoor plants. The floor is clean and tidy.

A suitcase is placed in a busy airport waiting lounge, with a background featuring travelers waiting and a display screen showing boarding information.

The smart bracelet, set against an outdoor sports environment, features a backdrop of mountains and blue skies. Surrounded by grass and natural landscapes, it creates a lifestyle that is close to nature.

Orange juice, placed on an outdoor picnic blanket, with a background of lush green grass in a park. Next to the orange juice are sandwiches and fresh fruits, surrounded by happy families.

Sneakers placed in a vast mountainous and leafy outdoor environment, with distant peaks and clear lakes in the background. The sneakers tread on a firm dirt path surrounded by lush greenery.

The chopstick and bowl storage box is placed on the supermarket shelf, with its vibrant packaging attracting customers' attention. Surrounding shelves are filled with various kitchen supplies.

A blender is placed in an environment that emphasizes healthy eating, with a background of nutritious ingredients and sunlight streaming through the window.

Figure S8. Some cases in our user study.

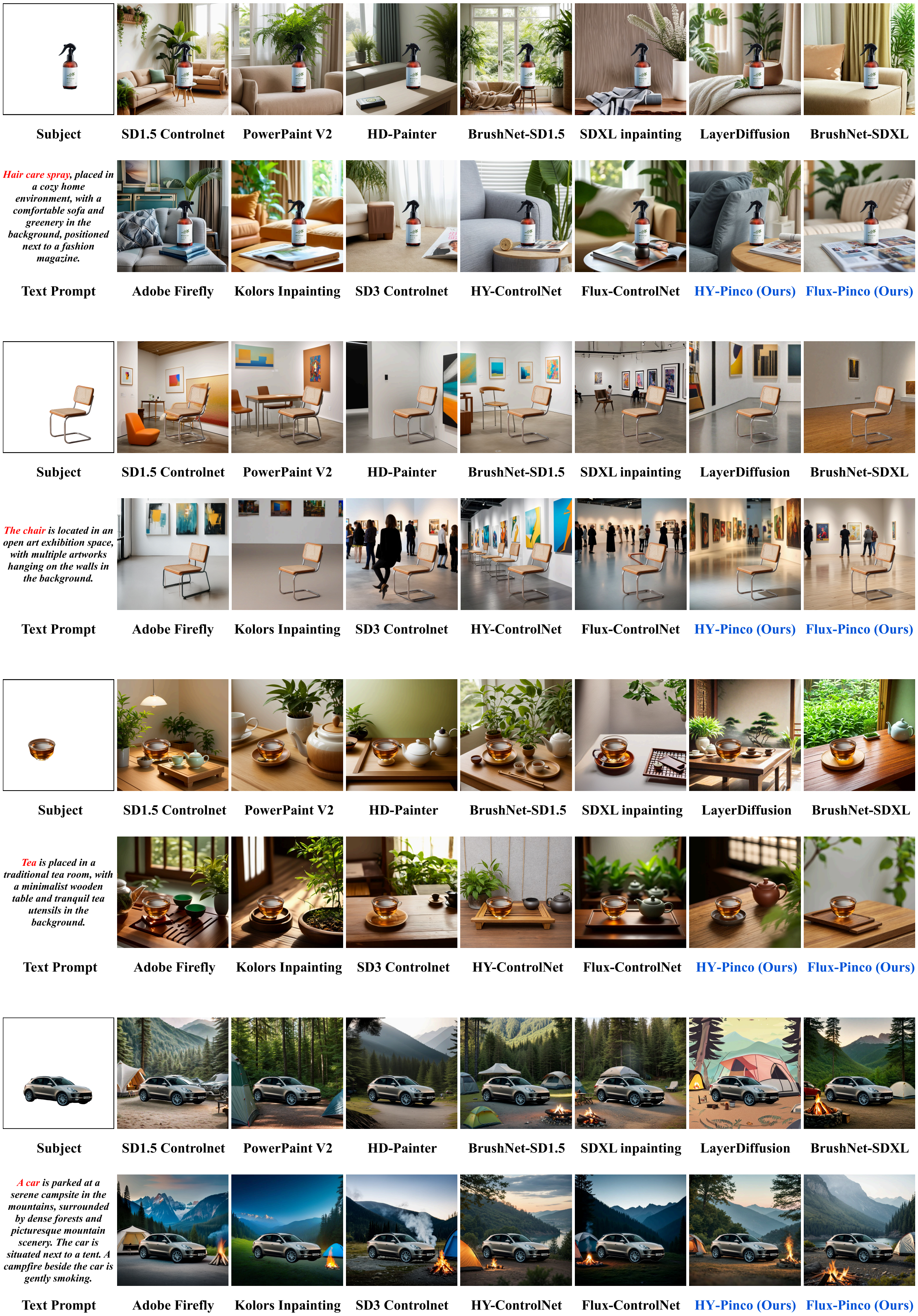
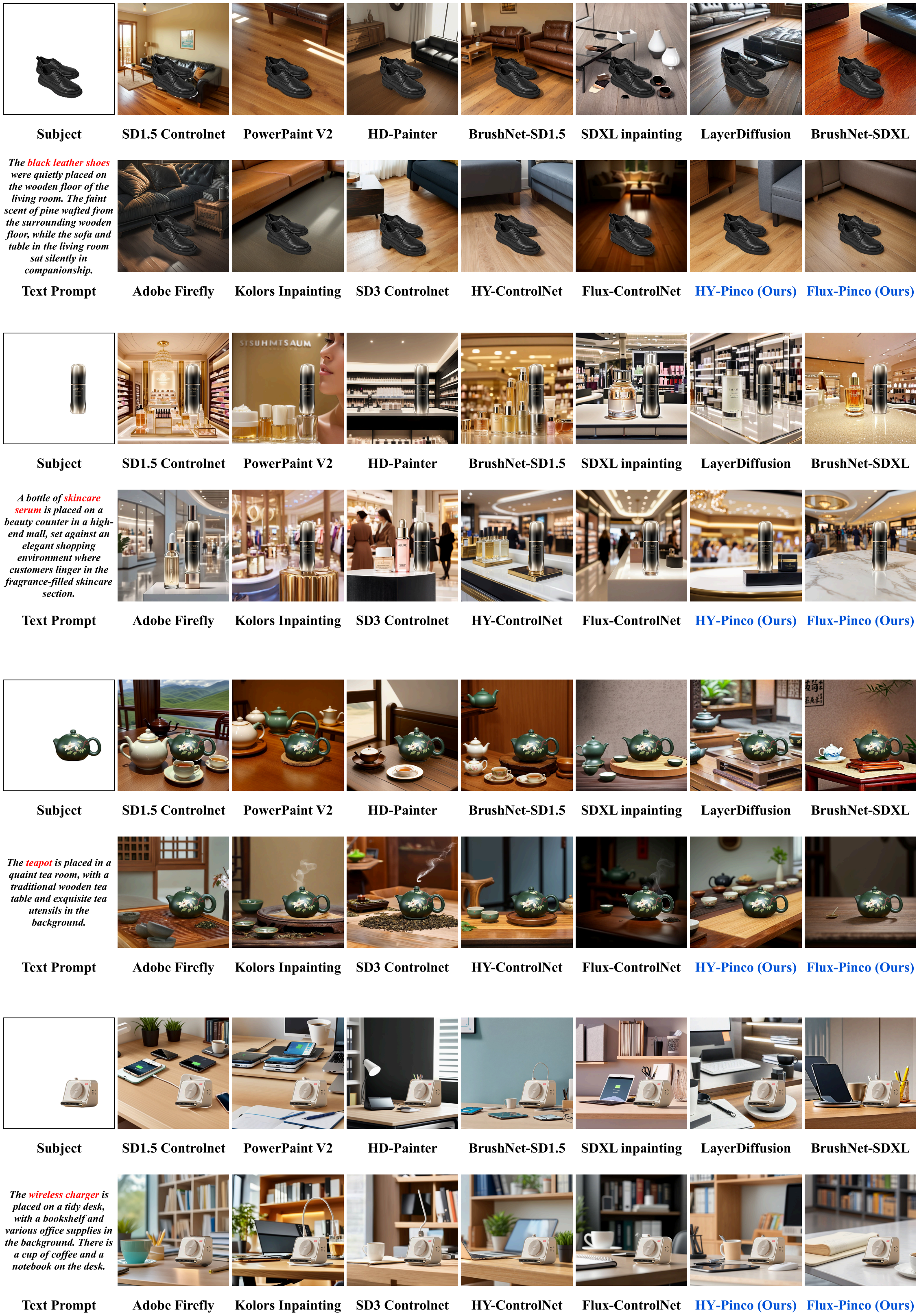


Figure S9. More qualitative comparisons between our Pinco and the state-of-the-art methods.



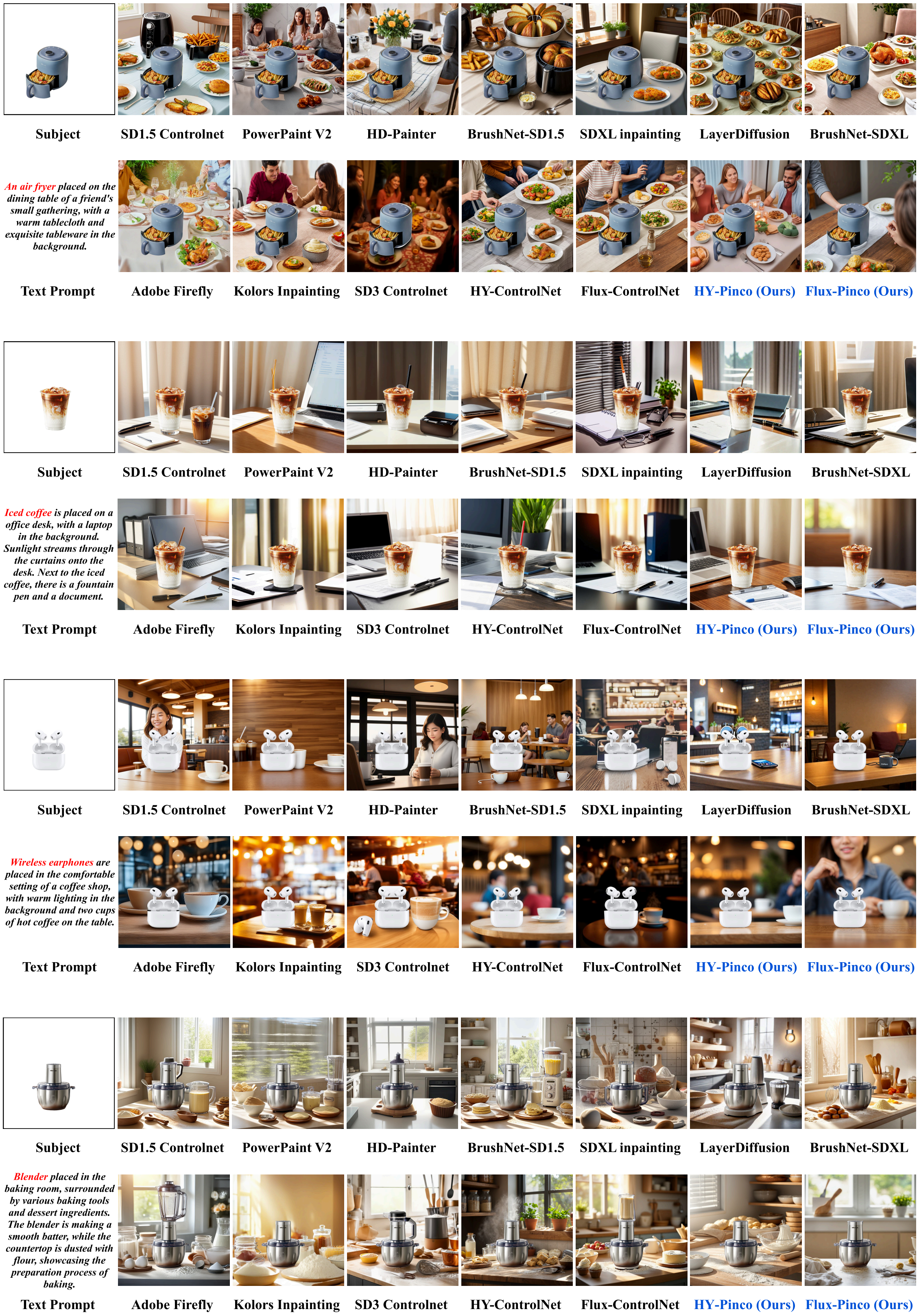
The black leather shoes were quietly placed on the wooden floor of the living room. The faint scent of pine wafted from the surrounding wooden floor, while the sofa and table in the living room sat silently in companionship.

A bottle of skincare serum is placed on a beauty counter in a high-end mall, set against an elegant shopping environment where customers linger in the fragrance-filled skincare section.

The teapot is placed in a quaint tea room, with a traditional wooden tea table and exquisite tea utensils in the background.

The wireless charger is placed on a tidy desk, with a bookshelf and various office supplies in the background. There is a cup of coffee and a notebook on the desk.

Figure S10. More qualitative comparisons between our Pinco and the state-of-the-art methods.



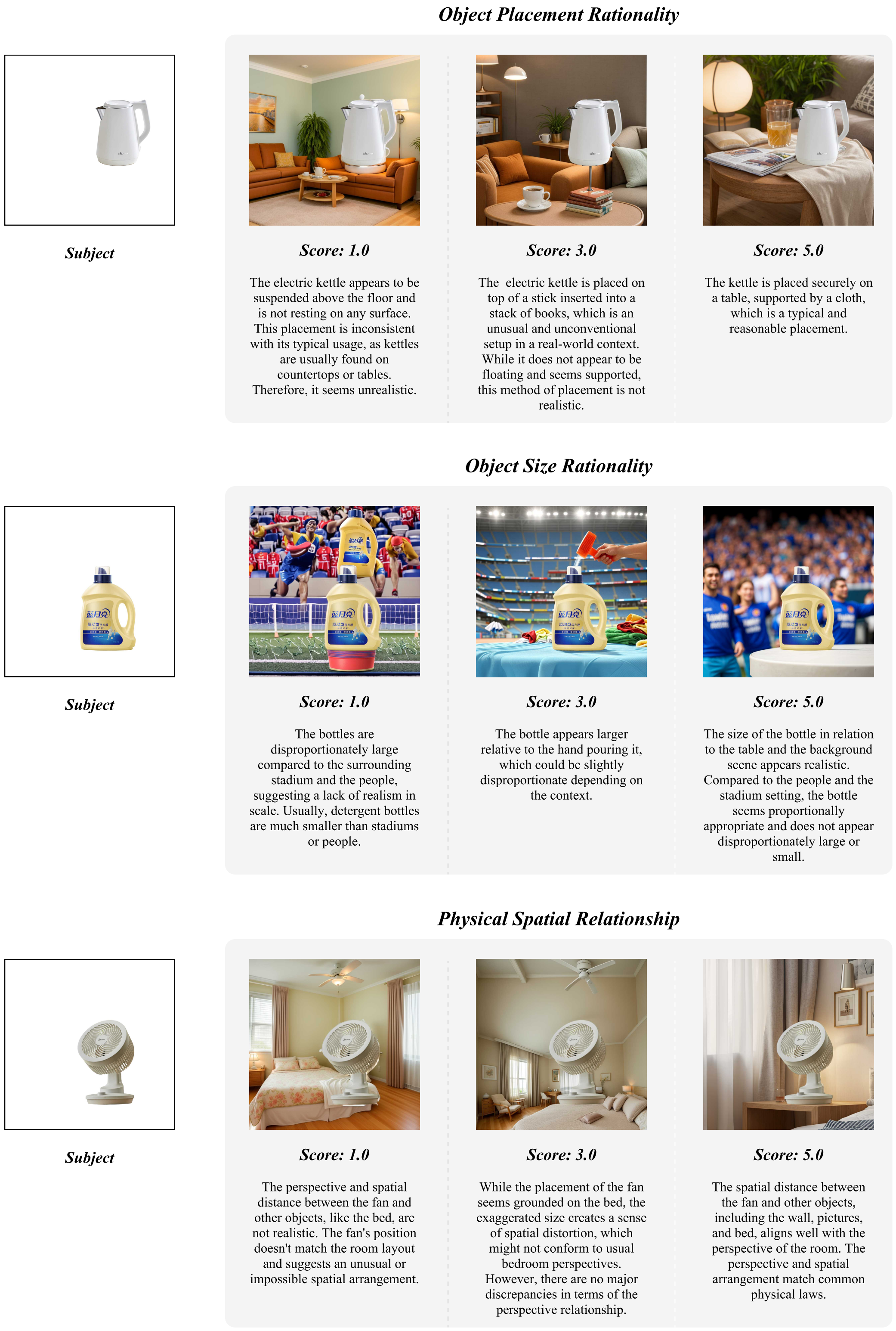
An air fryer placed on the dining table of a friend's small gathering, with a warm tablecloth and exquisite tableware in the background.

Iced coffee is placed on a office desk, with a laptop in the background. Sunlight streams through the curtains onto the desk. Next to the iced coffee, there is a fountain pen and a document.

Wireless earphones are placed in the comfortable setting of a coffee shop, with warm lighting in the background and two cups of hot coffee on the table.

Blender placed in the baking room, surrounded by various baking tools and dessert ingredients. The blender is making a smooth batter, while the countertop is dusted with flour, showcasing the preparation process of baking.

Figure S11. More qualitative comparisons between our Pinco and the state-of-the-art methods.



Object Placement Rationality

Subject

Score: 1.0

The electric kettle appears to be suspended above the floor and is not resting on any surface. This placement is inconsistent with its typical usage, as kettles are usually found on countertops or tables. Therefore, it seems unrealistic.

Score: 3.0

The electric kettle is placed on top of a stick inserted into a stack of books, which is an unusual and unconventional setup in a real-world context. While it does not appear to be floating and seems supported, this method of placement is not realistic.

Score: 5.0

The kettle is placed securely on a table, supported by a cloth, which is a typical and reasonable placement.

Object Size Rationality

Subject

Score: 1.0

The bottles are disproportionately large compared to the surrounding stadium and the people, suggesting a lack of realism in scale. Usually, detergent bottles are much smaller than stadiums or people.

Score: 3.0

The bottle appears larger relative to the hand pouring it, which could be slightly disproportionate depending on the context.

Score: 5.0

The size of the bottle in relation to the table and the background scene appears realistic. Compared to the people and the stadium setting, the bottle seems proportionally appropriate and does not appear disproportionately large or small.

Physical Spatial Relationship

Subject

Score: 1.0

The perspective and spatial distance between the fan and other objects, like the bed, are not realistic. The fan's position doesn't match the room layout and suggests an unusual or impossible spatial arrangement.

Score: 3.0

While the placement of the fan seems grounded on the bed, the exaggerated size creates a sense of spatial distortion, which might not conform to usual bedroom perspectives. However, there are no major discrepancies in terms of the perspective relationship.

Score: 5.0

The spatial distance between the fan and other objects, including the wall, pictures, and bed, aligns well with the perspective of the room. The perspective and spatial arrangement match common physical laws.

Figure S12. GPT-4o rationality analysis results.