A$^{\text{T}}$A: Adaptive Transformation Agent for Text-Guided Subject-Position Variable Background Inpainting (Supplementary Material)

Yizhe Tang[1*] Zhimin Sun[1,2*] Yuzhen Du[1,2] Ran Yi[1†] Guangben Lu[2]
Teng Hu[1] Luying Li[2] Lizhuang Ma[1] Fangyuan Zou[2]
[1]Shanghai Jiao Tong University [2]Tencent

S1. Overview

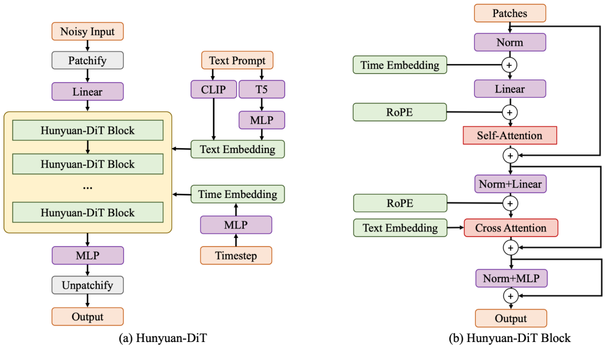
In this supplementary material, we mainly present the following contents:

- More technical details of our A$^{\text{T}}$A model structure (Sec. S2);
- More technical details of the evaluation metrics (Sec. S3);
- More qualitative comparison for the subject-position variable inpainting task (Sec. S4);
- Qualitative comparison for the subject-position fixed inpainting task with user-provided layout (Sec. S5);
- More details of the user study (Sec. S6);
- More results of A$^{\text{T}}$A with different aspect ratios (Sec. S7);
- Visualization and analysis of the attention map (Sec. S8);
- More analysis of the RDT module (Sec. S9);
- Prospect of future works (Sec. S10).

S2. More Details of Model Structure

In this section, we provide a more detailed description of the model structure. We adopt the architecture Hunyuan-DiT-g/2 [4] as our base model, which has 40 blocks and an embedding dimension of $1,408$. Our A$^{\text{T}}$A model consists of 4 modules: Feature extraction, Reverse displacement transform, Feature fusion, and Diffusion denoising (refer to the main paper Fig. 2 for overall architecture).

Feature Extraction. We use a tiny Swin-Transformer [7] backbone as the feature extractor for the subject image, and pre-process the input with an 8×8 window size. This process yields a set of multi-scale feature maps, denoted as $\{C_I^1, \cdots, C_I^N\}$, with $N = 4$ and the dimensions of C_I^1 being $C \times H \times W$. Following each Swin-Transformer [7] stage, the number of channels is doubled compared to the previous stage, while the spatial dimensions (height and width) are halved, resulting in a new feature map of $2C \times H/2 \times W/2$ dimension. Since the 4 subject feature maps will be injected into the subject cross-attention module (after the displacement transform), which requires the injected feature to have a unified dimension, we use a convolutional layer to adjust the features, mapping them to the same dimension for future injection.



(a) Hunyuan-DiT (b) Hunyuan-DiT Block

Figure S1. The model structure of Hunyuan-DiT [4] (the base model of our A$^{\text{T}}$A).

Reverse Displacement Transform. Given that our foundational model is based on a DiT framework, we opt for the DiT block [9] enhanced with Adaptive Layer Norm (AdaLN) to serve as the architecture for the PosAgent block $\mathcal{A}_i$. The PosAgent block takes the output from the Swin-transformer Φ as its input, incorporating time embedding t, text embedding C_T, and positional switch embedding C_P on the conditional side, fusing them with the input via a Layer Normalization technique. The output from each of these blocks is then utilized to perform a reverse transformation on the feature map $\{C_I^1, \cdots, C_I^N\}$ initially extracted by the Swin-transformer Φ, effectively restructuring it in a reverse manner. *I.e.*, the transformation parameters predicted by the first PosAgent block $\mathcal{A}_1$ are used to transform the last (deepest) feature map C_I^N, while the last PosAgent block $\mathcal{A}_N$ will transform the first (shallowest) feature map C_I^1.

Feature Fusion. We employ a pre-trained cross-attention mechanism as our subject cross-attention module, which is initialized with identical weights to the base model's cross-

[*] Equal contribution. {tangyizhe, zhimin.sun}@sjtu.edu.cn.
[†] Corresponding author. ranyi@sjtu.edu.cn.

attention module. Subsequently, the output from this attention module is combined with the output from the original self-attention module using a trainable tanh weight. This setup allows the model to dynamically adjust the influence of these two attention modules depending on the specific requirements of various Hunyuan-DiT blocks.

Diffusion Denoising. After all features are fused, A$^{\mathrm{T}}$A performs a T-step denoising process as Hunyuan-DiT (structure shown in Fig. S1) and obtains the inpainted image.

S3. Evaluation Metrics

In this section, we provide a more detailed description of the evaluation metrics. We quantitatively evaluate the performance of the model from the following perspectives: Image Quality, Extension Ratio, Text-alignment, Multi-subject Rate, and Position Rationality.

1) Image Quality: We calculate the Fréchet Inception Distance (FID) [3] score on the MSCOCO [5] dataset to evaluate the quality of generated images.

2) Extension Ratio: To evaluate the subject extension ratio, we adopt the OER [2] metric which calculates the consistency between the foreground subject mask of the generated image against the ground truth subject mask. Specifically, we first use BiRefNet [11] to segment the generated image and obtain an accurate mask M of the foreground subject. For subject-position fixed background inpainting methods, since the subject's position is expected to be the same position in the original image, the subject mask of the original image $M_o = 1 - m$ serves as the ground truth subject mask. Then the OER score can be computed as follows:

$$OER = \frac{\sum \mathrm{ReLU}(M - M_o)}{\sum M_o}, \tag{S1}$$

where ReLU is the activation function. The smaller the OER score, the better subject consistency is achieved by the inpainting model. Since our A$^{\mathrm{T}}$A can adaptively determine a suitable position and size of the subject, and generate an inpainted image with the subject in the new position, we utilize FlorenceV2 [10] to detect a bounding box of the subject, and rescale the original subject mask M_o to fit the detected bounding box as the ground truth subject mask M_o'. The OER score for our A$^{\mathrm{T}}$A is:

$$OER = \frac{\sum \mathrm{ReLU}(M - M_o')}{\sum M_o'}. \tag{S2}$$

3) Text-alignment: To measure the text-image alignment, following Imagen3 [1], we choose VQA [6] score which evaluates the alignment between an image and a text prompt by using a visual-question-answering model to answer simple yes-or-no questions about the image content.

4) Multi-subject Rate: In the text-guided background inpainting task, sometimes the model will repeatedly draw the given foreground subject since it cannot recognize it in the prompt. In this case, the generated result will contain multiple similar subjects, which may not be what the user wants. As a result, we adopt FlorenceV2 [10] to detect and count the number of the subject in the generated result given the subject name. Then we calculate the ratio that the number of the detected subject is greater than the desired number of the subject as the multi-subject rate.

5) Postion Rationality: We leverage GPT-4o to evaluate each image based on placement, size, and spatial relationships. Each image is scored with a maximum score of 100, and the average score determines the rationality.

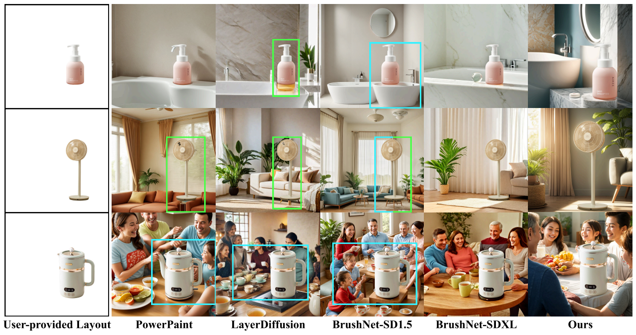


Figure S2. Comparison between our A$^{\mathrm{T}}$A, and existing methods using user-provided subject layout. The comparison methods exhibit unsuitable relative sizes and positional relationships between foreground and background; while our method achieves harmonious positional relationships.

S4. More Comparisons on Subject-Position Variable Inpainting

Fig. S4 presents additional qualitative comparisons against various methods on the Subject-Position Variable Inpainting task. Consistent with the results in the main paper, when combining previous inpainting methods with LayoutGPT for subject-position variable inpainting, they often generate results with layouts that are either illogical or have incorrect proportions. Even when a suitable layout is produced, previous inpainting methods fail to grasp subject positioning, causing background-text conflicts or missing objects. In contrast, our method A$^{\mathrm{T}}$A generates high-quality inpainted results with a harmonious positional relationship between the subject and the inpainted background, as well as good text-background alignment.

Furthermore, to eliminate the influence of LayoutGPT and facilitate fairer comparison, for the methods combined with LayoutGPT in main paper Fig. 5 and Fig. S4, we replace LayoutGPT with the user-provided subject layout. As shown in Fig. S2, the comparison methods exhibit unsuitable relative sizes and positional relationships between foreground and background.

S5. Comparisons on Subject-Position Fixed Inpainting with User-provided Layout

As mentioned in the main paper, our method is capable of both subject-position variable and subject-position fixed inpainting, which can be flexibly switched by setting the position switch embedding. To evaluate the performance on the Subject-Position Fixed Inpainting task, we have undertaken an exhaustive qualitative analysis, with the same user-provided subject layout for both comparison methods and our method. As depicted in Fig. S5, our approach $\mathrm{A^TA}$ demonstrates its effectiveness by generating satisfactory inpainting results when the position of the subject is fixed. The previous inpainting methods suffer from problems including missing certain objects in the background (not well aligned with text), subject expansion, multiple subjects (of inaccurate number), and inappropriate subject size or subject mispositioned. In contrast, $\mathrm{A^TA}$ generates high-quality results with minimal occurrences of subjects expanding beyond the boundaries or multiple subjects, and achieves a good text-background alignment. Moreover, our method excels at creating a more harmonious visual relationship between the generated background and the subject.

S6. Details of User Study

We conduct a user study to assess the rationality of the subject position and overall quality of the inpainted results. In the user study, we invited 31 participants majoring in computer science to conduct the experiment, and each participant received 40 sets of test questions. Fig. S6 presents some sample sets in the user study. Each set of test questions consists of 2 inpainted images: one generated by $\mathrm{A^TA}$ and another from a different method, along with the source image and the text prompt. Each set of test images are shuffled to ensure that the questionnaire is blindly evaluated by the participants. Participants are asked to choose the better image based on *the rationality of the subject position* and overall image quality. Then we calculate the average preference between $\mathrm{A^TA}$ and the other 4 compared methods, and the results are shown in Fig. 6 in the main paper, where our $\mathrm{A^TA}$ receives the most preference from the participants, demonstrating our superior position rationality.

S7. More Results of Different Aspect Ratios

We conduct extensive comparisons for images with Different Aspect Ratios. As illustrated in Figures S7, S8, S9, $\mathrm{A^TA}$ demonstrates its versatility by producing high-quality image outputs across a range of aspect ratios.

S8. Visualization & Analysis of Attention Map

To provide a more intuitive assessment of the position switch's performance, we perform an attention map visualization experiment. This experiment involves comparing the model's attention maps (of the subject cross-attention layers) with the position switch embedding C_P activated (subject position change enabled) and deactivated (subject position fixed). Fig. S3 illustrates the changes of the attention distribution as the denoising process progresses. When the position switch is enabled (enabling adaptive subject position change), $C_P = E_v$, the model's attention gradually shifts towards the subject, effectively adapting to the position offset. In contrast, when the position switch is disabled (fixing the subject position to that of the original image), $C_P = E_o$, the model's attention remains concentrated on the subject region in the original image, without any position offset. Without subject-position adaptation, the model can generate a well-inpainting image, but its subject-background position relationship may not be as harmonious as when the position switch is activated.

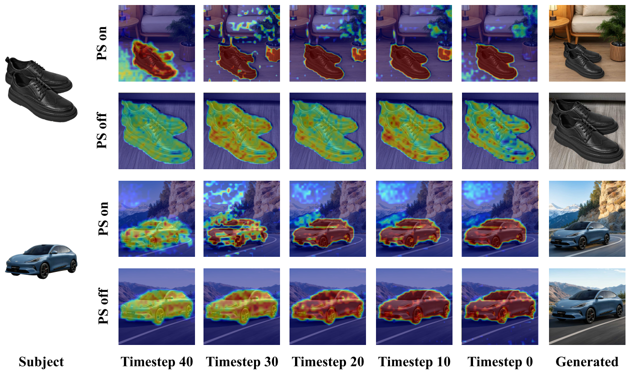


Figure S3. The visualization of attention maps of our model with position switch (PS) on and off. As the denoising process progresses, the model with the position switch on gradually focuses its attention towards the subject, effectively adapting to the position offset; while when the position switch is off, the model's attention remains concentrated on the subject region in the original image, without any position offset.

S9. Analysis of RDT module

In Sec. 4.3 and Sec. S2, we introduce the pipeline and design details of the proposed Reverse Displacement Transform module. In this section, we analyze the necessity of the RDT module. The RDT module is designed to predict a suitable position for the subject, where the text information is injected through the condition channel, which is the focus of Subject-Position Variable Inpainting task. Without RDT module, the extracted feature only contains the subject appearance information, which will lead to the generated subject being in a totally random position or even with distortions during the subjection fusion stage. We further evaluate the additional cost brought by the RDT module, including the time cost during training stage and inference stage, as shown in Tab. S1. Following Pinco [8], we also leverage

Table S1. Analysis of RDT module, where we test the time cost during training stage and inference stage and the GPT-4o score.

Methods	training(s/epoch)↓	inference(s/image)↓	rationality by GPT4o↑
w/o RDT	976.5	8.41	86.2
w/ RDT	1044.5	8.94	92.8

GPT-4o to evaluate each image based on the rationality of the subject's position (with maximum score 100). From Tab. S1, we can see that the RDT module has markedly improved the position rationality score (86.2→92.8) while bringing marginal time cost to the overall training (6.5%) and inference (5.9%) pipelines.

S10. Prospect

In our proposed "Text-Guided Subject-Position Variable Background Inpainting task", we only focus on the position of one single subject, aiming to adaptively adjust the single subject position and generate a harmonious image. However, for the multi-subjects scenario, it would become more complex since the relative positioning and hierarchical constraints should be taken into consideration. Different from Subject-Position Fixed Inpainting where the multi-subjects can be combined in one image as the input, for Subject-Position Variable Inpainting, the multi-subjects should be adaptively adjusted separately, but with the constraint of relative position to align with the text prompt. Also, the conflict of multi-subjects' positions should be avoided where the overlap of multi-subjects might happen. In conclusion, the multi-subjects scenario is quite a meaningful but difficult direction, and our next step will consider multi-subject adaptive positioning and might adopt an additional relative position rationality module for evaluation and constraint.

References

[1] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024. 2

[2] Binghui Chen, Chongyang Zhong, Wangmeng Xiang, Yifeng Geng, and Xuansong Xie. Virtualmodel: Generating object-id-retentive human-object interaction image by diffusion model for e-commerce marketing. *arXiv preprint arXiv:2405.09985*, 2024. 2

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[4] 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

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[6] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025. 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[8] Guangben Lu, Yuzhen Du, Zhimin Sun, Ran Yi, Yifan Qi, Yizhe Tang, Tianyi Wang, Lizhuang Ma, and Fangyuan Zou. Pinco: Position-induced consistent adapter for diffusion transformer in foreground-conditioned inpainting. *arXiv preprint arXiv:2412.03812*, 2024. 3

[9] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[10] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 2

[11] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *arXiv preprint arXiv:2401.03407*, 2024. 2

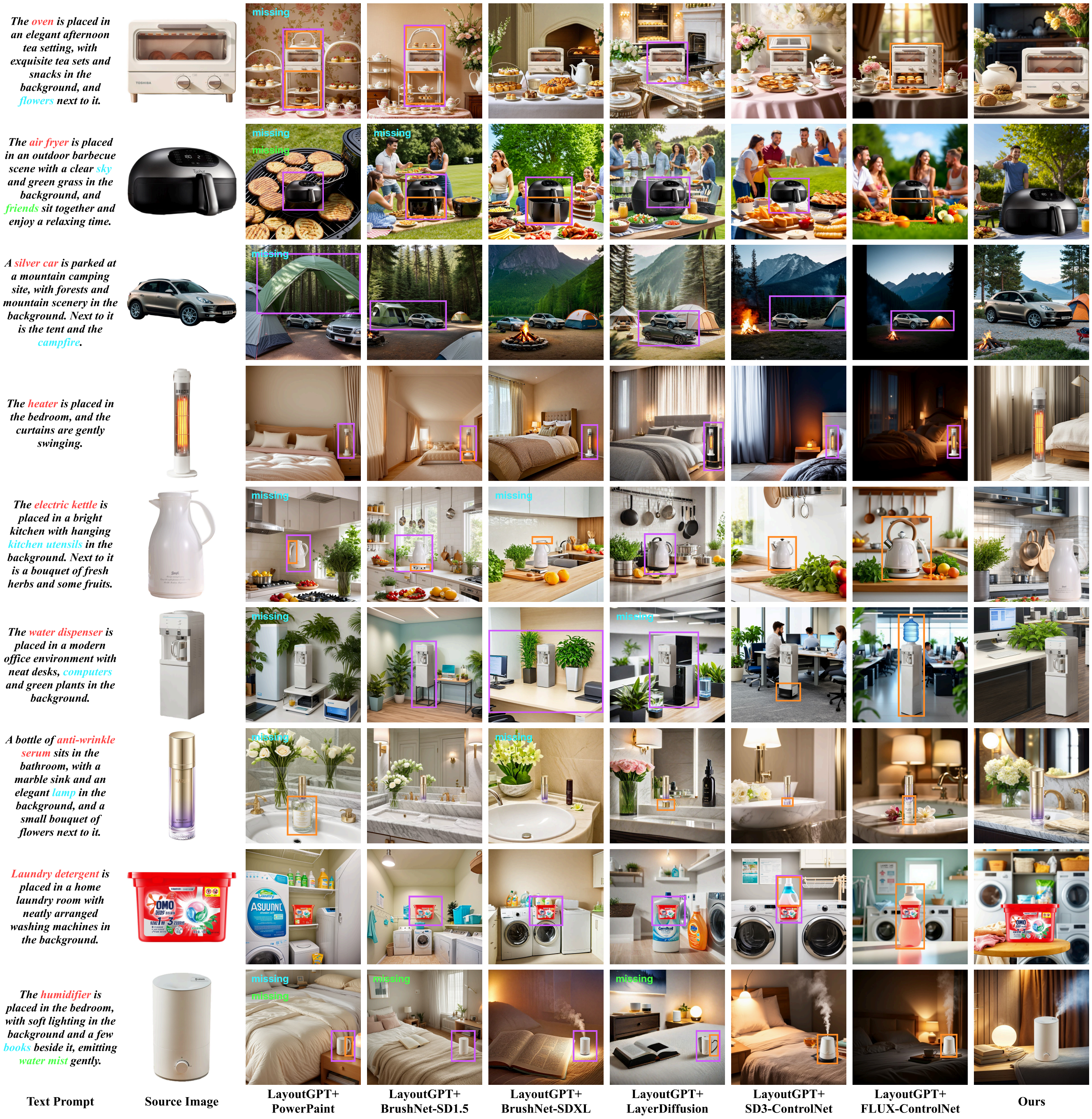
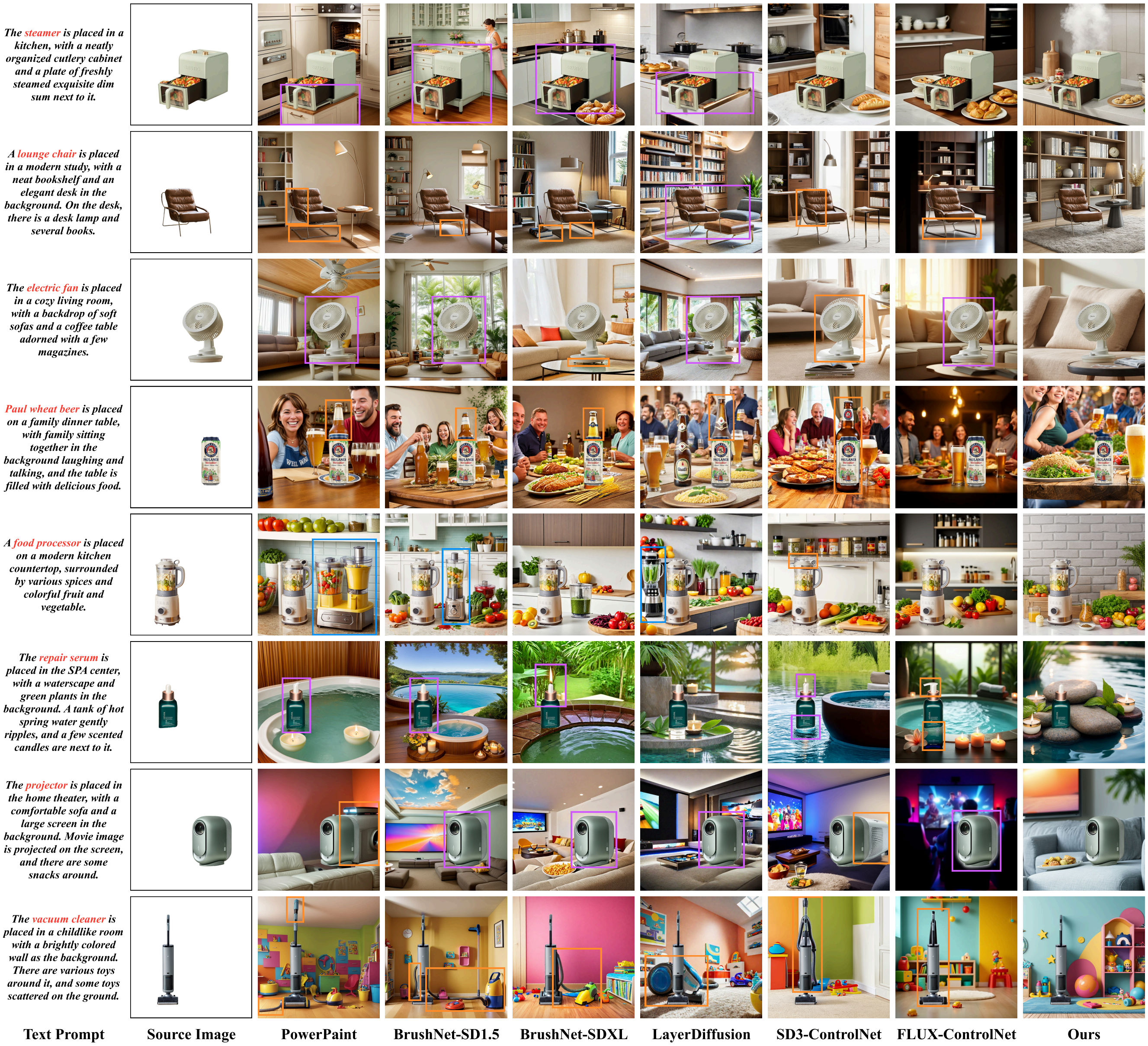

The oven is placed in an elegant afternoon tea setting, with exquisite tea sets and snacks in the background, and flowers next to it.
The air fryer is placed in an outdoor barbecue scene with a clear sky and green grass in the background, and friends sit together and enjoy a relaxing time.
A silver car is parked at a mountain camping site, with forests and mountain scenery in the background. Next to it is the tent and the campfire.
The heater is placed in the bedroom, and the curtains are gently swinging.
The electric kettle is placed in a bright kitchen with hanging kitchen utensils in the background. Next to it is a bouquet of fresh herbs and some fruits.
The water dispenser is placed in a modern office environment with neat desks, computers and green plants in the background.
A bottle of anti-wrinkle serum sits in the bathroom, with a marble sink and an elegant lamp in the background, and a small bouquet of flowers next to it.
Laundry detergent is placed in a home laundry room with neatly arranged washing machines in the background.
The humidifier is placed in the bedroom, with soft lighting in the background and a few books beside it, emitting water mist gently.
missing
Text Prompt
Source Image
LayoutGPT+ PowerPaint
LayoutGPT+ BrushNet-SD1.5
LayoutGPT+ BrushNet-SDXL
LayoutGPT+ LayerDiffusion
LayoutGPT+ SD3-ControlNet
LayoutGPT+ FLUX-ControlNet
Ours

Figure S4. More Qualitative results for the *Subject-Position Variable* Inpainting task. We highlight the unreasonable extension parts with orange boxes and the unreasonable layouts with purple boxes, and label the missing objects with corresponding colors. Please zoom in for more details.



The steamer is placed in a kitchen, with a neatly organized cutlery cabinet and a plate of freshly steamed exquisite dim sum next to it.

A lounge chair is placed in a modern study, with a neat bookshelf and an elegant desk in the background. On the desk, there is a desk lamp and several books.

The electric fan is placed in a cozy living room, with a backdrop of soft sofas and a coffee table adorned with a few magazines.

Paul wheat beer is placed on a family dinner table, with family sitting together in the background laughing and talking, and the table is filled with delicious food.

A food processor is placed on a modern kitchen countertop, surrounded by various spices and colorful fruit and vegetable.

The repair serum is placed in the SPA center, with a waterscape and green plants in the background. A tank of hot spring water gently ripples, and a few scented candles are next to it.

The projector is placed in the home theater, with a comfortable sofa and a large screen in the background. Movie image is projected on the screen, and there are some snacks around.

The vacuum cleaner is placed in a childlike room with a brightly colored wall as the background. There are various toys around it, and some toys scattered on the ground.

Figure S5. More Qualitative results for the *Subject-Position Fixed* Inpainting task. We highlight the unreasonable extension parts with orange boxes, the unreasonable layouts with purple boxes and the multi-subjects with blue boxes. Please zoom in for more details.

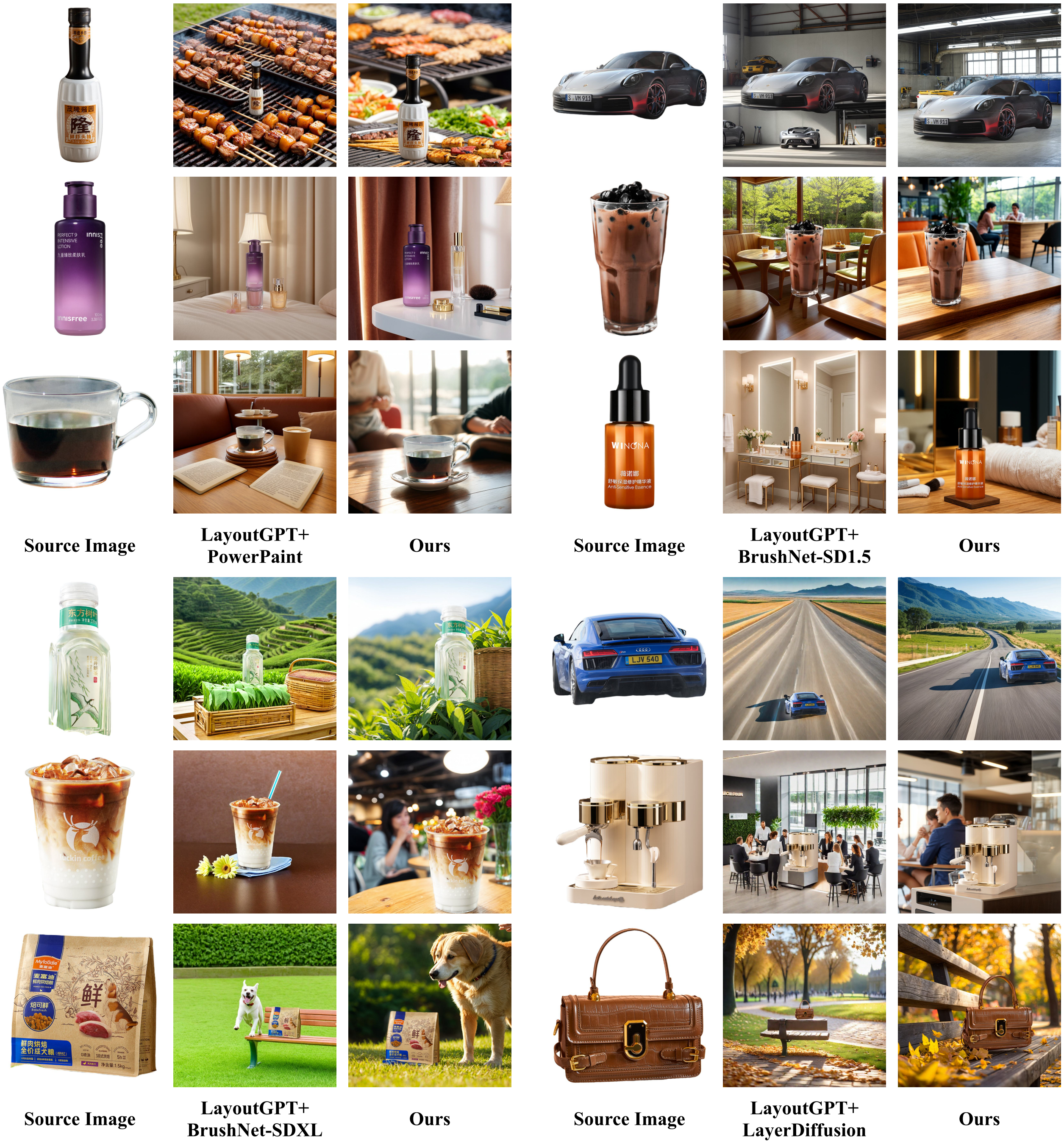


Figure S6. Some example sets used in our user study. Here we show the names of methods for ease of visual comparison. However, in the actual user study, the order of the images was shuffled and participants did not know the names of the methods.

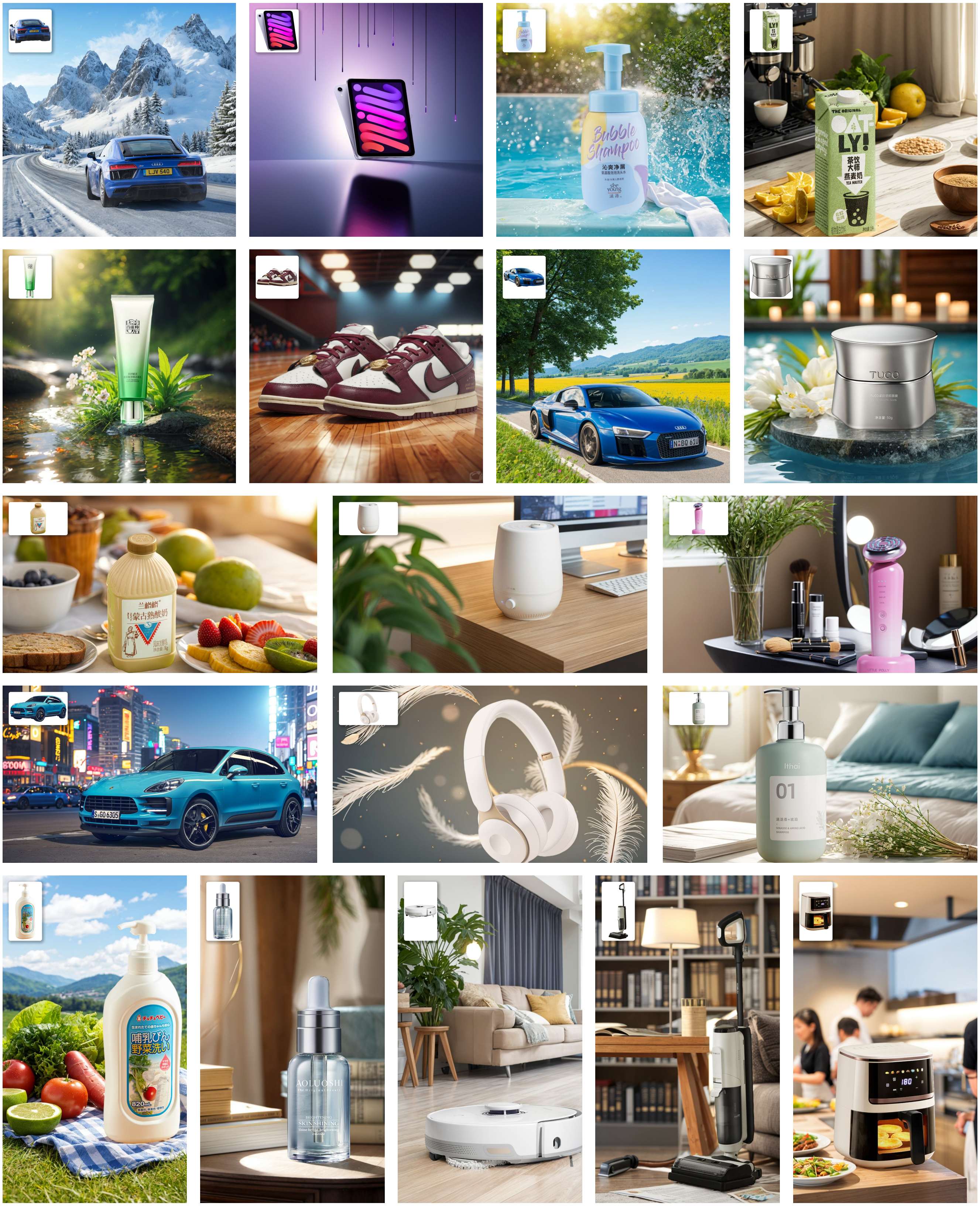


Figure S7. Our A^{T}A can generate high-quality inpainted results across different aspect ratios.

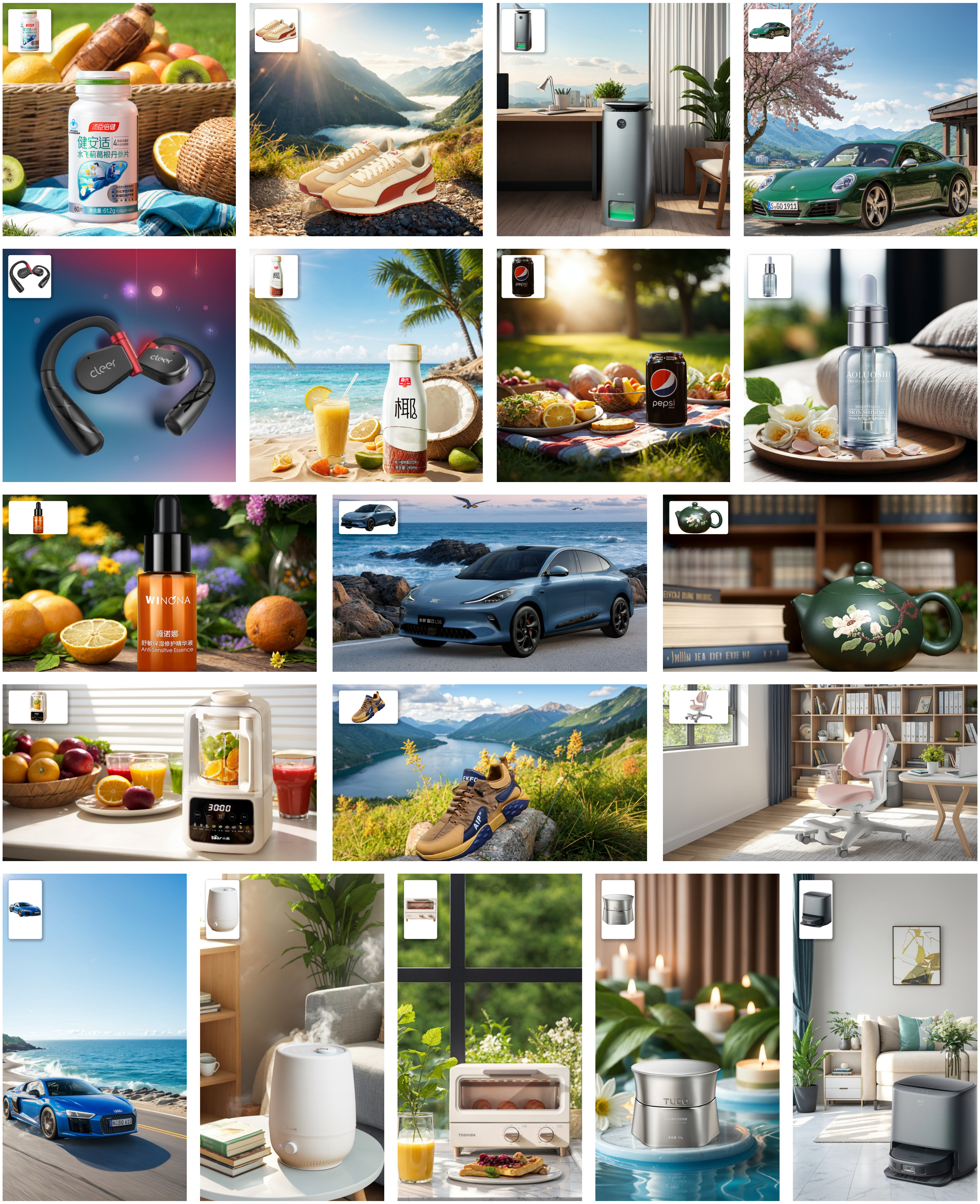


Figure S8. Our A^{T}A can generate high-quality inpainted results across different aspect ratios.

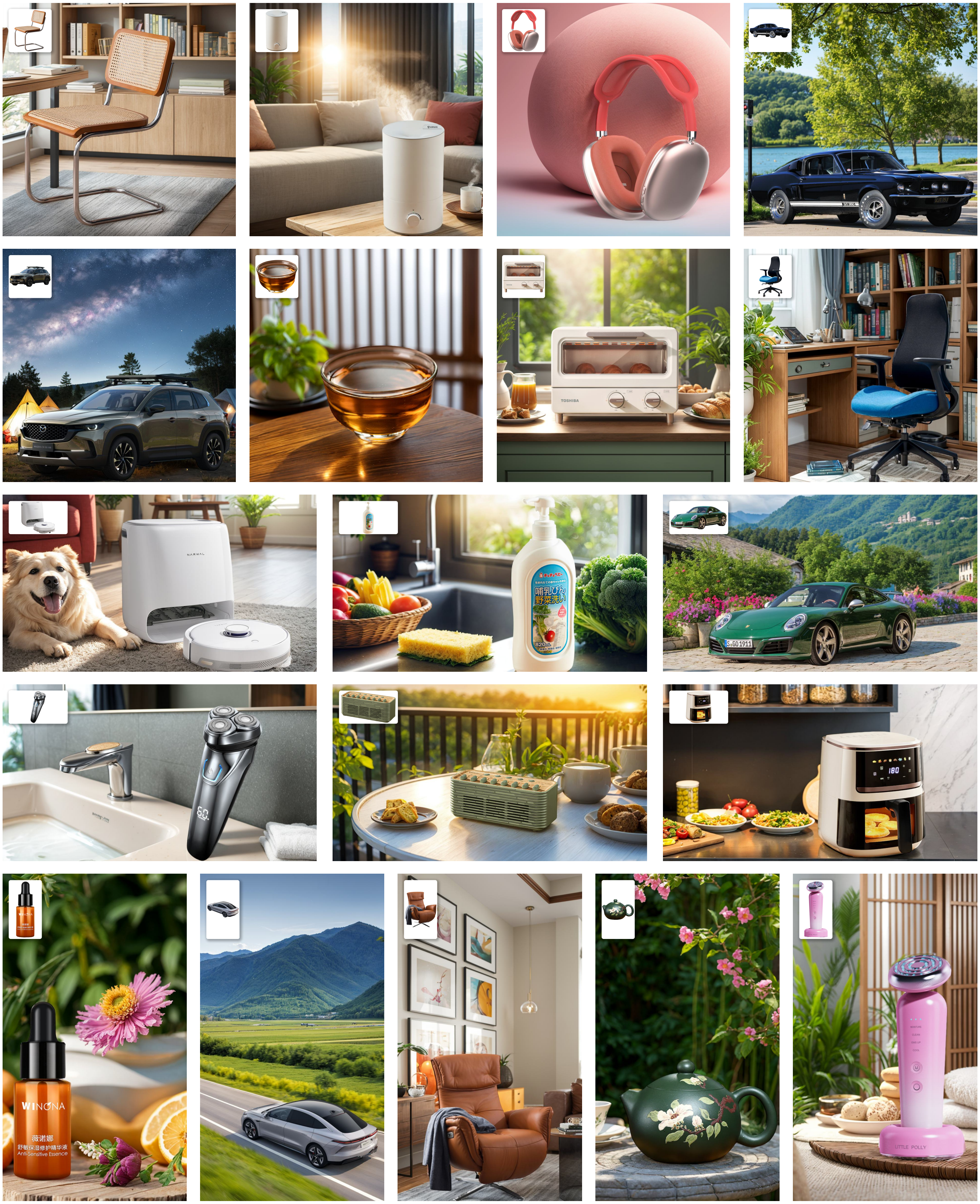


Figure S9. Our A^{T}A can generate high-quality inpainted results across different aspect ratios.