Equilibrated Diffusion: Frequency-aware Textual Embedding for Equilibrated Image Customization Supplementary Material

Liyuan Ma
Westlake University
Hangzhou, China
maliyuan@westlake.edu.cn

Xueji Fang
Zhejiang University
Westlake University
Hangzhou, China
fangxueji@zju.edu.cn

Guo-Jun Qi*
Westlake University
Hangzhou, China
guojunq@gmail.com

1 METHOD DETAILS

1.1 Implementation Details

Our method is based on the publicly available Stable Diffusion 1.4, which is consistent with other comparing methods. During the optimization process, our method involves learnable parameters comprising the projection matrices of Key and Value in the cross-attention component of the Stable Diffusion's UNet, the projection matrices of Key and Value within the reference attention of residual reference attention (RRA), and two learnable text embeddings corresponding to the low-frequency and high-frequency inputs in Frequency-aware Decoupled Textual Embedding (FDTE).

Regarding the prompt format in our method, we employed <S^*> to signify the learnable identifier and <C> to denote the category of the subject. For instance, in phrases like "a photo of <S^*> <C> on the beach," <C> denotes categories such as toy or cat. Notably, the text embedding of <S^*> was derived from the frequency embedding in the main manuscript. During inference, the embedding's value of <S^*> was obtained by summing low-frequency and high-frequency embeddings. In the training phase, the text embedding of <S^*> was chosen with specified probabilities from a composite of embeddings associated with low-frequency, high-frequency, and their cumulative representation.

1.2 Prompt Templates

To examine how our method performs across different prompts and to affirm its capability in stylized descriptions, we devised non-stylized and stylized prompt templates for generating custom images. These templates were subsequently utilized for inference and metrics calculation. The detailed prompt templates were provided as follows.

1.2.1 Unstylized Prompt Templates. Considering the contrast between the live and non-live reference subject images, we have utilized different prompts to sample unstylized images. These prompts have been adapted from DreamMatcher [4] and consist of both challenging and non-challenging variations. However, we have combined them in our experiment, resulting in a total of 55 prompts. For cases involving non-live subjects, the prompts are outlined as follows:

- a <S^*> <C> in the jungle
- a <S^*> <C> in the snow
- a <S^*> <C> on the beach
- a <S^*> <C> on a cobblestone street
- a <S^*> <C> on top of pink fabric
- a <S^*> <C> on top of a wooden floor
- a <S^*> <C> with a city in the background
- a <S^*> <C> with a mountain in the background
- a <S^*> <C> with a blue house in the background
- a <S^*> <C> on top of a purple rug in a forest
- a <S^*> <C> with a wheat field in the background
- a <S^*> <C> with a tree and autumn leaves in the background
- a <S^*> <C> with the Eiffel Tower in the background
- a <S^*> <C> floating on top of water
- a <S^*> <C> floating in an ocean of milk
- a <S^*> <C> on top of green grass with sunflowers around it
- a <S^*> <C> on top of a mirror
- a <S^*> <C> on top of the sidewalk in a crowded street
- a <S^*> <C> on top of a dirt road
- a <S^*> <C> on top of a white rug
- a red <S^*> <C>
- a purple <S^*> <C>
- a shiny <S^*> <C>
- a wet <S^*> <C>
- a <S^*> <C> with Japanese modern city street in the background
- a <S^*> <C> with a landscape from the Moon
- a <S^*> <C> among the skyscrapers in New York city
- a <S^*> <C> with a beautiful sunset
- a <S^*> <C> in a movie theater
- a <S^*> <C> in a luxurious interior living room
- a <S^*> <C> in a dream of a distant galaxy
- a <S^*> <C> floating in a pond
- a <S^*> <C> bouncing on a trampoline
- a <S^*> <C> on a wooden dock by a lake
- a <S^*> <C> in a grassy park with a bench
- a <S^*> <C> on a brick pathway in a garden
- a <S^*> <C> at the edge of a swimming pool
- a <S^*> <C> on a stone wall in the countryside
- a <S^*> <C> in a schoolyard playground
- a <S^*> <C> on a sandy beach near the dunes
- a <S^*> <C> at a picnic spot with a checkered blanket
- a <S^*> <C> partially covered by sand in the desert
- a <S^*> <C> nestled among rocks
- a <S^*> <C> inside a box
- a <S^*> <C> inside a closet
- a <S^*> <C> between two chairs

[*] Guo-Jun Qi is the corresponding author.

- a <S^*> <C> inside a basket
- a <S^*> <C> wearing a top hat
- a <S^*> <C> wearing a scarf
- a <S^*> <C> in an astronaut outfit
- a <S^*> <C> wearing bowtie
- a <S^*> <C> seen from the top
- a <S^*> <C> seen from the bottom
- a <S^*> <C> seen from the back
- a <S^*> <C> seen from the side

For cases of live subjects, the prompts are listed as follows:

- a <S^*> <C> in the jungle
- a <S^*> <C> in the snow
- a <S^*> <C> on the beach
- a <S^*> <C> on a cobblestone street
- a <S^*> <C> on top of pink fabric
- a <S^*> <C> on top of a wooden floor
- a <S^*> <C> with a city in the background
- a <S^*> <C> with a mountain in the background
- a <S^*> <C> with a blue house in the background
- a <S^*> <C> on top of a purple rug in a forest
- a <S^*> <C> wearing a red hat
- a <S^*> <C> wearing a santa hat
- a <S^*> <C> wearing a rainbow scarf
- a <S^*> <C> wearing a black top hat and a monocle
- a <S^*> <C> in a chef outfit
- a <S^*> <C> in a firefighter outfit
- a <S^*> <C> in a police outfit
- a <S^*> <C> wearing pink glasses
- a <S^*> <C> wearing a yellow shirt
- a <S^*> <C> in a purple wizard outfit
- a red <S^*> <C>
- a purple <S^*> <C>
- a shiny <S^*> <C>
- a wet <S^*> <C>
- a <S^*> <C> with Japanese modern city street in the background
- a <S^*> <C> with a landscape from the Moon
- a <S^*> <C> among the skyscrapers in New York city
- a <S^*> <C> with a beautiful sunset
- a <S^*> <C> in a movie theater
- a <S^*> <C> in a luxurious interior living room
- a <S^*> <C> in a dream of a distant galaxy
- a <S^*> <C> jumping over a cascading waterfall
- a <S^*> <C> swinging from the trees in a rainforest
- a <S^*> <C> jumping over a fence
- a <S^*> <C> climbing a tree
- a <S^*> <C> swinging on a swing
- a <S^*> <C> diving into a pool
- a <S^*> <C> running across a meadow
- a <S^*> <C> hopping on stepping stones
- a <S^*> <C> leaping across rooftops in a city
- a <S^*> <C> swimming upstream in a river
- a <S^*> <C> submerged halfway in a crystal-clear lake
- a <S^*> <C> looking out from a tent
- a <S^*> <C> camouflaged in tall grass
- a <S^*> <C> behind a cloud of smoke
- a <S^*> <C> inside a cave entrance
- a <S^*> <C> wearing a top hat
- a <S^*> <C> wearing a backpack
- a <S^*> <C> wearing a crown
- a <S^*> <C> in an astronaut outfit
- a <S^*> <C> wearing a bowtie
- a <S^*> <C> seen from the top
- a <S^*> <C> seen from the bottom
- a <S^*> <C> seen from the back
- a <S^*> <C> seen from the side

Here <S^*> <C> would be replaced by a rare token, instance name, or a combination of them, depending on the specific subject.

1.2.2 Stylized Prompt Templates. In order to gauge the adherence of our model to the stylized image customization, we utilized LLM [5] to produce 22 stylized prompt templates as follows.

- <S^*> <C> oil painting ghibli inspired
- Painting of <S^*> <C> at a beach by artist claude monet
- Georgia O'Keeffe style <S^*> <C> painting
- watercolor painting of a <S^*> <C>
- Painting of a <S^*> <C> by artist Claude Monet
- Georgia O'Keeffe style desert <S^*> <C> painting
- <S^*> <C> impressionistic painting of a lively cityscape
- Photorealistic still life painting of a bowl of <S^*> <C>
- A surreal <S^*> <C> sculpture in bronze
- abstract painting of <S^*> <C>
- Pointillism-inspired <S^*> <C> artwork
- A vibrant Van Gogh-inspired <S^*> <C> painting
- Abstract sculpture of a dancing <S^*> <C> by Henry Moore
- A whimsical <S^*> <C> watercolor illustration
- Hyperrealistic portrait of a <S^*> <C> by artist Chuck Close
- A minimalist <S^*> <C> painting
- Impressionistic <S^*> <C> inspired by Monet
- Surreal dream-like artwork featuring a <S^*> <C>
- A mosaic representation of a <S^*> <C>
- Pointillism-inspired artwork of a <S^*> <C> on sofa
- Cubist-style portrait of a <S^*> <C> by Picasso
- Sketch of a <S^*> <C>

2 ADDITIONAL EXPERIMENT ANALYSIS

2.1 Comparison with FreeStyle

FreeStyle [2] emerges as a training-free style transfer method adept at transforming reference images into novel renditions based on provided textual style descriptions. At the core of FreeStyle is a sophisticated architecture composed of a dual-stream encoder and a single-stream decoder. The proposed architecture achieves distinct encodings for both the content image and style textual prompts, facilitating their seamless integration during the decoding phase. Notably, FreeStyle introduces two scaling factors, **s** and **b**, alongside an additional truncation parameter, **n**, to modulate the feature maps generated by diverse encoders, thereby finely adjusting the balance between style infusion and content fidelity. s governs the strength of style guidance from textual prompts, whereas b and n regulate the degree of content preservation within the content image. A higher s intensifies the stylization effect in the outcomes. Furthermore, with increasing values of b and n, the consistency with the original image amplifies.

To better highlight our method's capability in stylizing reference images under stylized text descriptions, FreeStyle serves as a highly effective benchmark. Comparison with FreeStyle demonstrates our method's capacity to trade off between stylizing specific subjects and maintaining concept consistency.

As depicted in fig. 1, the style transfer effect of FreeStyle is highly sensitive to hyperparameters and not universally effective. When employing the default settings of the original manuscript (s=1, b=2 or 3, n=320), FreeStyle often struggles to generate results that align with the stylized textual description. Achieving the desired style representation becomes feasible after selecting appropriate hyperparameters tailored to specific concepts, although maintaining the original semantics of the content still poses challenges. This observation highlights the limitations of FreeStyle in style transfer. Conversely, our generation results as shown in fig. 1 demonstrate the ability to stylize the content reasonably while maintaining subject consistency, indicating the advantage of our method over training-free approach in customized content stylization.

2.2 User Study

We additionally conducted a user study to compare our approach with previous works. Specifically, we randomly chose 20 cases, with each case paired with different prompts. Half of the prompts were selected from the unstylized prompt templates, while the other half belonged to stylized prompts. For each case, we compared Textual Inversion [1], DreamBooth [6], DreamMatcher [4], Custom Diffusion [3], ELITE [7], resulting in a total of 6 generated images along with ours. These images were shuffled and presented to users for ranking based on their perceived quality of image customization.

To provide a better interactive interface, we designed an online questionnaire form hosted on *wjx.cn* as shown in fig. 2. The first 10 questions focused on text alignment comparison, while the remaining 10 questions were related to image alignment. The given question for the text alignment section was "Sort the following options according to the similarity of the text descriptions,". For image alignment comparison, participants are instructed to rank options according to the instruction like "Sort the following options according to the similarity of the given reference image.". Finally, we collected 42 questionnaires from users for 20 comparative sorting questions, resulting in a total of 840 answers and 12.6k comparisons (15 comparisons per sorting).

We presented the numerical results of user study in fig. 3 and table 1. These results demonstrate that our approach achieves the highest subjective evaluation scores in terms of text and image consistency of generating results and outperforms previous methods by a large margin, thus showcasing the superiority of our method.

2.3 More Visual Results

We presented more visual results of different customization methods in fig. 4, fig. 5, fig. 6 and fig. 7. These results further verify the conclusions presented in our main manuscript, demonstrating the superior adaptability of our method compared to others across various complex prompts and reference images, as well as its ability to maintain consistency with textual descriptions and reference subjects.

3 LIMITATIONS

While RRA and FDTE improve subject consistency with reference and textual alignment, they do face challenges. As shown in fig. 8, the generated result lacks some structural elements compared with the reference, like the missing pole of the car, indicating RRA struggles to restore the subject accurately facing complex subject structures. Subject overfitting is also an issue where the generated result does not fully align with the stylized text even with enhanced prompt alignment by FDTE, as seen with the red color leakage in the right part of fig. 8.

4 SOCIAL IMPACTS

Our model is capable of generating customized images based on user-provided images, adhering to given themes. It excels not only in placing people or objects in various scenes but also particularly shines in generating images in different artistic styles. This aspect raises concerns regarding potential issues such as portrait rights and copyright of artworks. We will implement content filtering on the generated results to ensure they do not pose threats to individuals' privacy and security. It is noteworthy that although our model leads in custom generation, the resulting images still exhibit slight deviations from natural images detectable by other models and generated by artificial intelligence. Thus, the spread of our model remains controllable and secure.

REFERENCES

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations.* https://openreview.net/forum?id=NAQvF08TcyG

[2] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, and Fanzhang Li. 2024. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636* (2024).

[3] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1931–1941.

[4] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. 2024. DreamMatcher: Appearance Matching Self-Attention for Semantically-Consistent Text-to-Image Personalization. *arXiv preprint arXiv:2402.09812* (2024).

[5] R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article* 2, 5 (2023).

[6] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22500–22510.

[7] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 15943–15953.

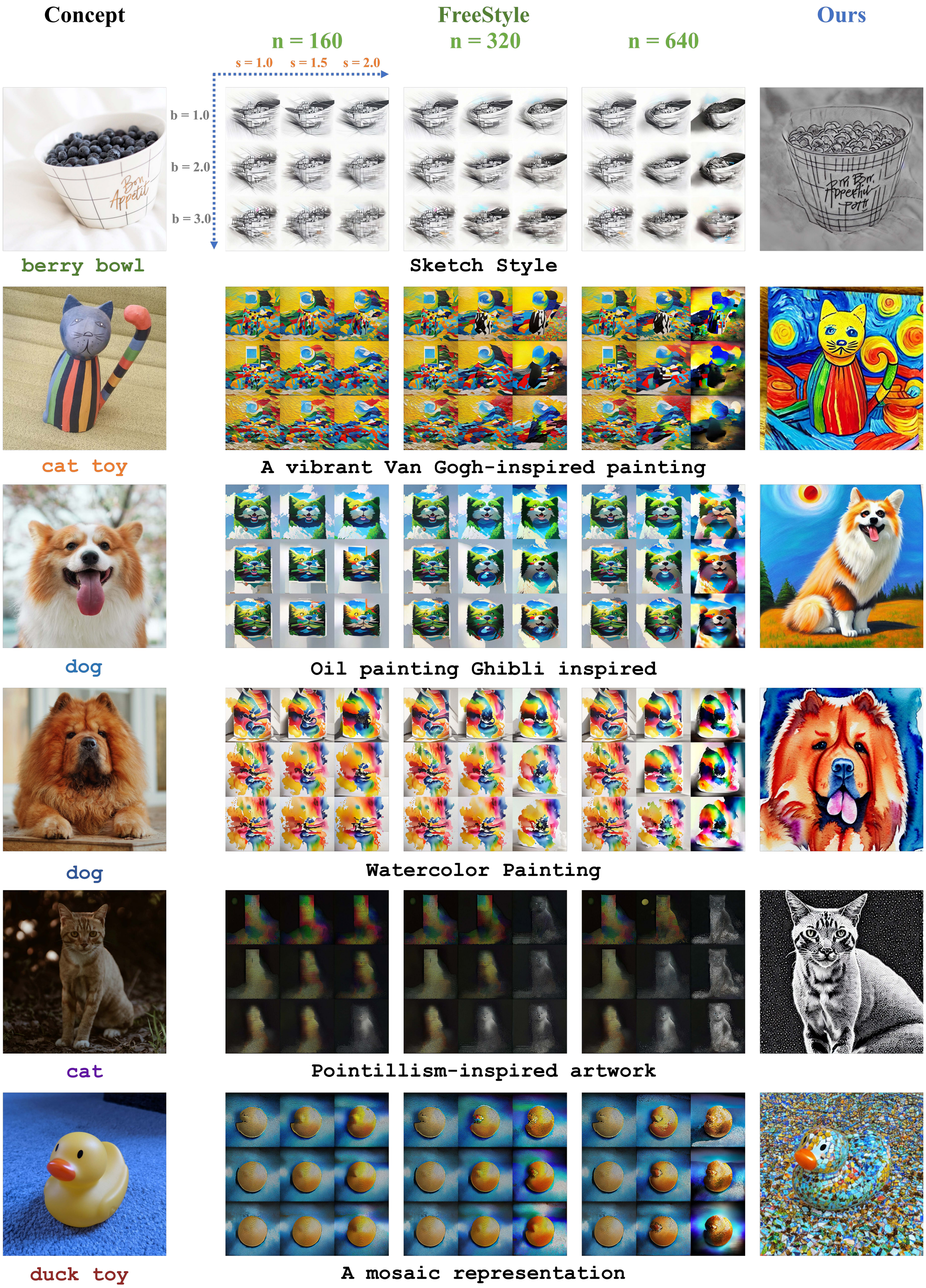
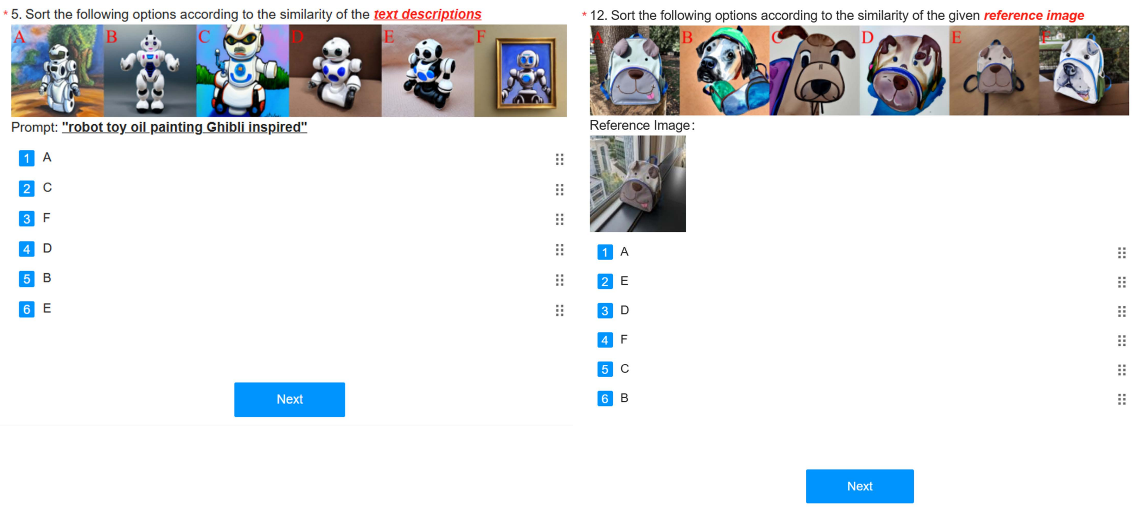


Figure 1: Visual comparison between training-free style-transfer method FreeStyle [2] with our approach. Please zoom in for better view.



* 5. Sort the following options according to the similarity of the *text descriptions*

Prompt: "robot toy oil painting Ghibli inspired"

1 A
2 C
3 F
4 D
5 B
6 E

Next

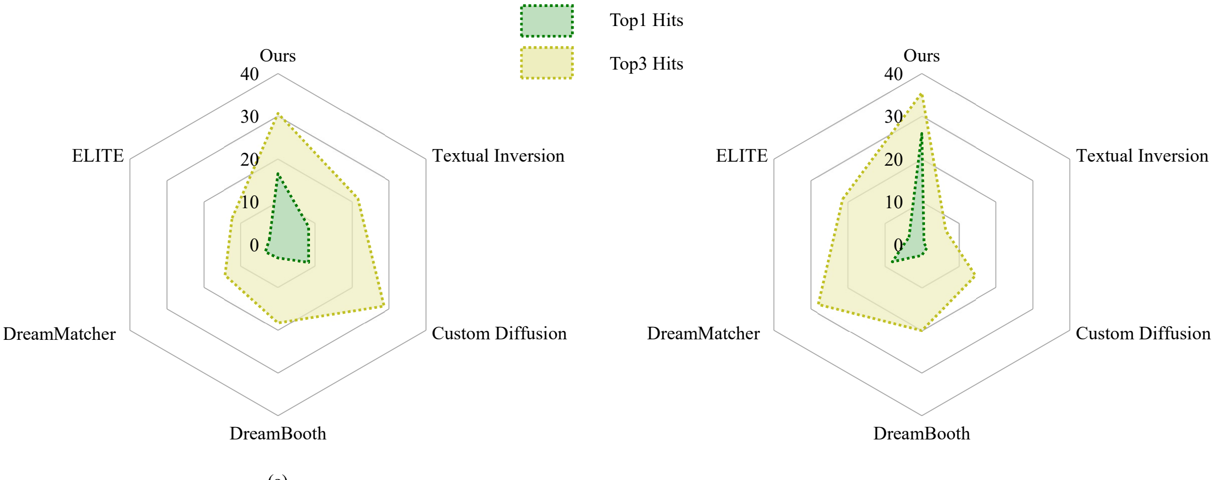
* 12. Sort the following options according to the similarity of the given *reference image*

Reference Image:

1 A
2 E
3 D
4 F
5 C
6 B

Next

Figure 2: User Study UI visualization. We designed various instructions for users to evaluate the effectiveness of text *(left)* alignment and image *(right)* alignment in customized image results.



(a) (b)

Figure 3: Comparison of Radar Chart Representations for Top1 and Top3 Hits in User Study. (a) Text alignment. (b) Image alignment. *Note: TopX Hits means the average number of times the model ranked as Top-X in total 42 questionnaires.*

Model	Text Alignment				Image Alignment			
	Top1 Hits	Top3 Hits	Top1 Hits Rates	Top3 Hits Rates	Top1 Hits	Top3 Hits	Top1 Hits Rates	Top3 Hits Rates
Textual Inversion [1]	8.3	21.6	19.76%	17.14%	0.6	6.5	1.43%	5.16%
Custom Diffusion [3]	8.3	28.7	19.76%	22.78%	1.4	14.5	3.33%	11.51%
DreamBooth [6]	3.1	18.3	7.38%	14.52%	2.5	20.1	5.95%	15.95%
DreamMatcher [4]	3.4	14.3	8.10%	11.35%	8	28	19.05%	22.22%
ELITE [7]	2.3	12.4	5.48%	9.84%	3.5	21.4	8.33%	16.98%
Ours	**16.6**	**30.7**	**39.52%**	**24.37%**	**26**	**35.5**	**61.90%**	**28.17%**

Table 1: Quantitative comparison of Top1 and Top3 Hits in User Study. The best and the second best results are boldfaced and underlined. *Note: TopX Hits Rates denotes the percentage of the average number of times model ranked as Top-X in total 42 questionnaires.*

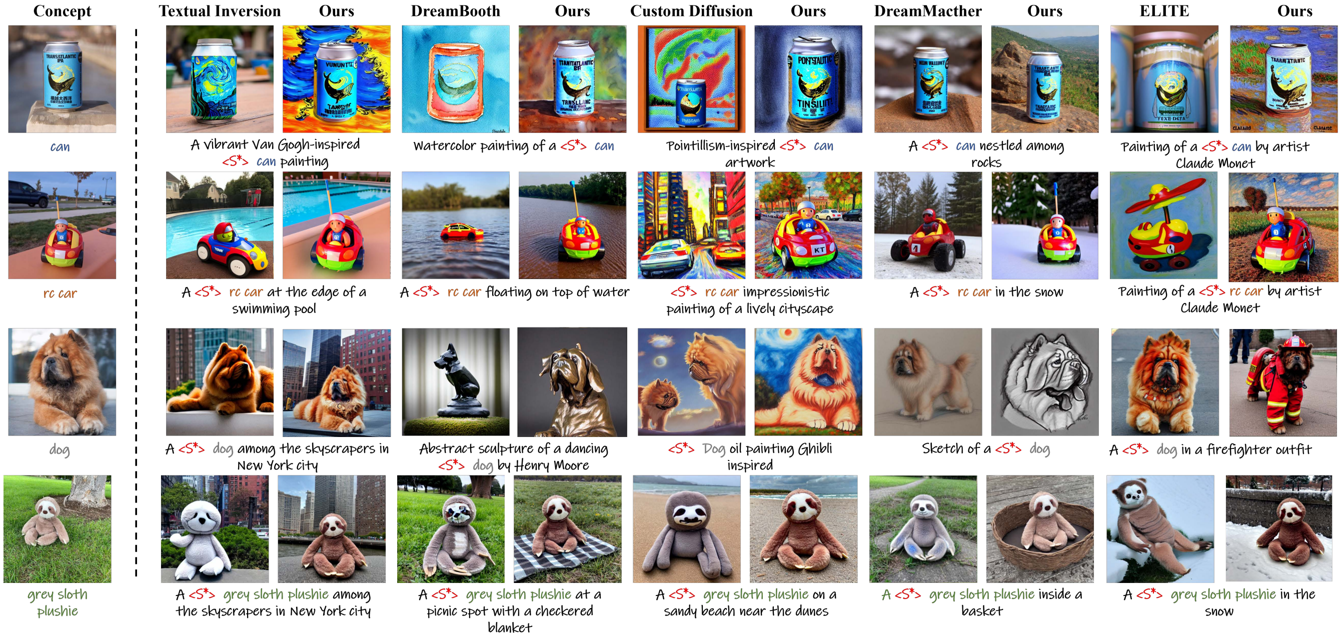

Concept
Textual Inversion
Ours
DreamBooth
Ours
Custom Diffusion
Ours
DreamMacther
Ours
ELITE
Ours
can
A vibrant Van Gogh-inspired <S*> can painting
Watercolor painting of a <S*> can
Pointillism-inspired <S*> can artwork
A <S*> can nestled among rocks
Painting of a <S*> can by artist Claude Monet
rc car
A <S*> rc car at the edge of a swimming pool
A <S*> rc car floating on top of water
<S*> rc car impressionistic painting of a lively cityscape
A <S*> rc car in the snow
Painting of a <S*> rc car by artist Claude Monet
dog
A <S*> dog among the skyscrapers in New York city
Abstract sculpture of a dancing <S*> dog by Henry Moore
<S*> Dog oil painting Ghibli inspired
Sketch of a <S*> dog
A <S*> dog in a firefighter outfit
grey sloth plushie
A <S*> grey sloth plushie among the skyscrapers in New York city
A <S*> grey sloth plushie at a picnic spot with a checkered blanket
A <S*> grey sloth plushie on a sandy beach near the dunes
A <S*> grey sloth plushie inside a basket
A <S*> grey sloth plushie in the snow

Figure 4: Qualitative visual results of comparing methods under varying prompts.

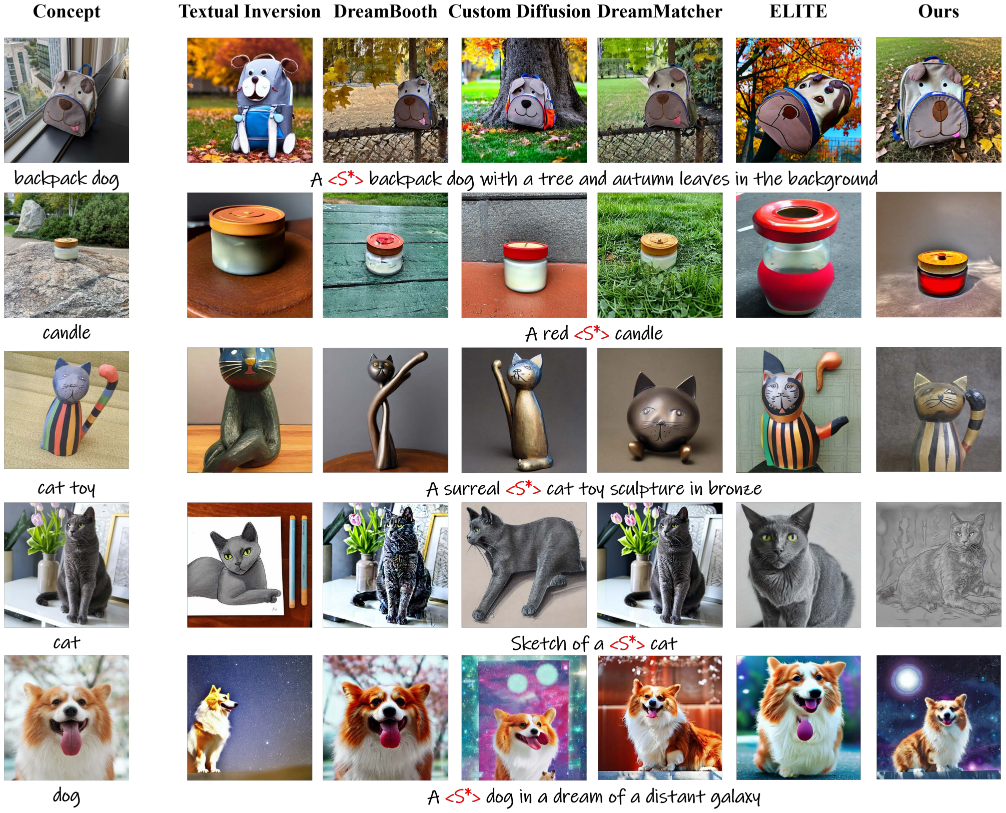

Concept
Textual Inversion
DreamBooth
Custom Diffusion
DreamMatcher
ELITE
Ours
backpack dog
A <S*> backpack dog with a tree and autumn leaves in the background
candle
A red <S*> candle
cat toy
A surreal <S*> cat toy sculpture in bronze
cat
Sketch of a <S*> cat
dog
A <S*> dog in a dream of a distant galaxy

Figure 5: Qualitative visual results of comparing methods under the same prompts.

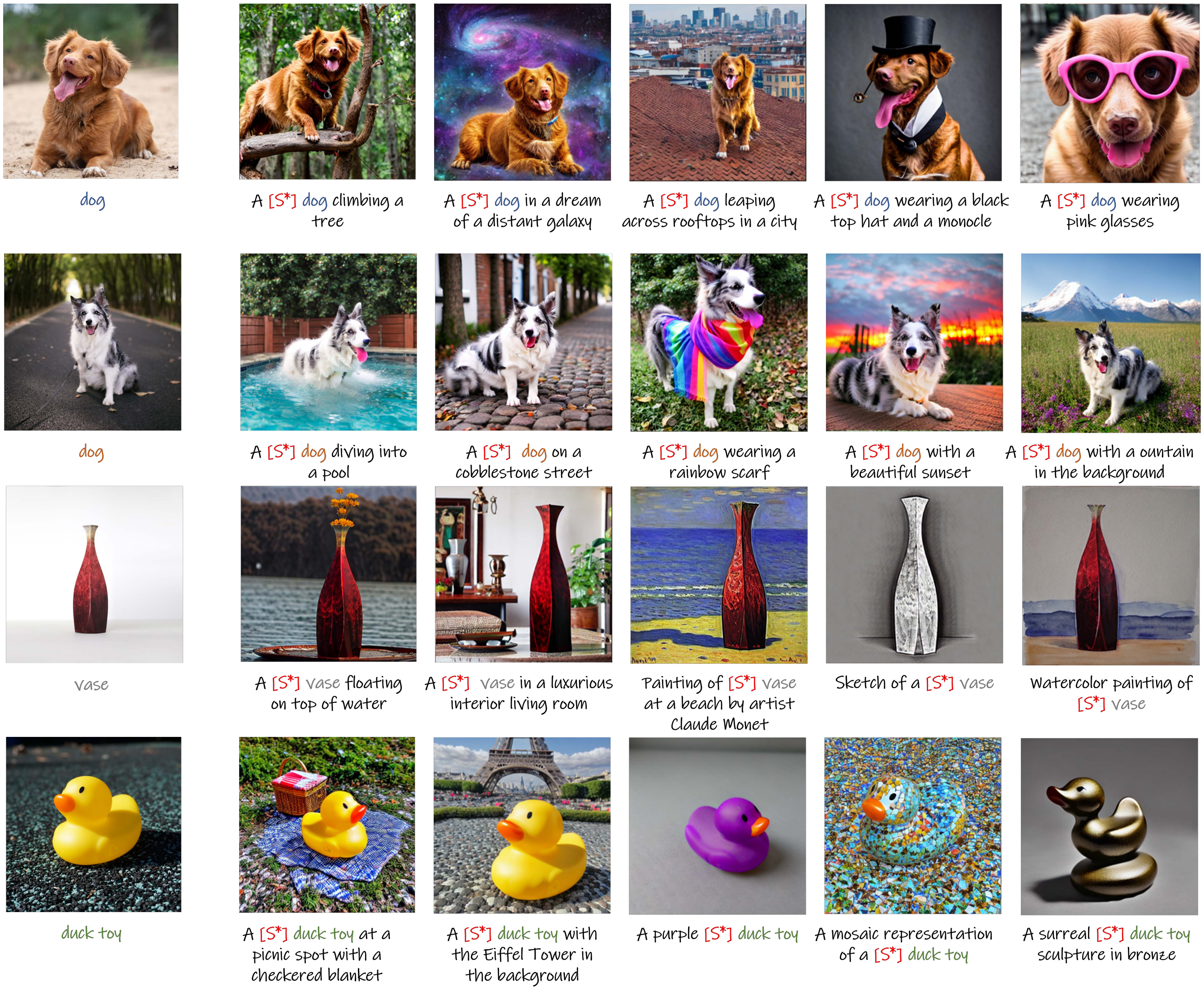
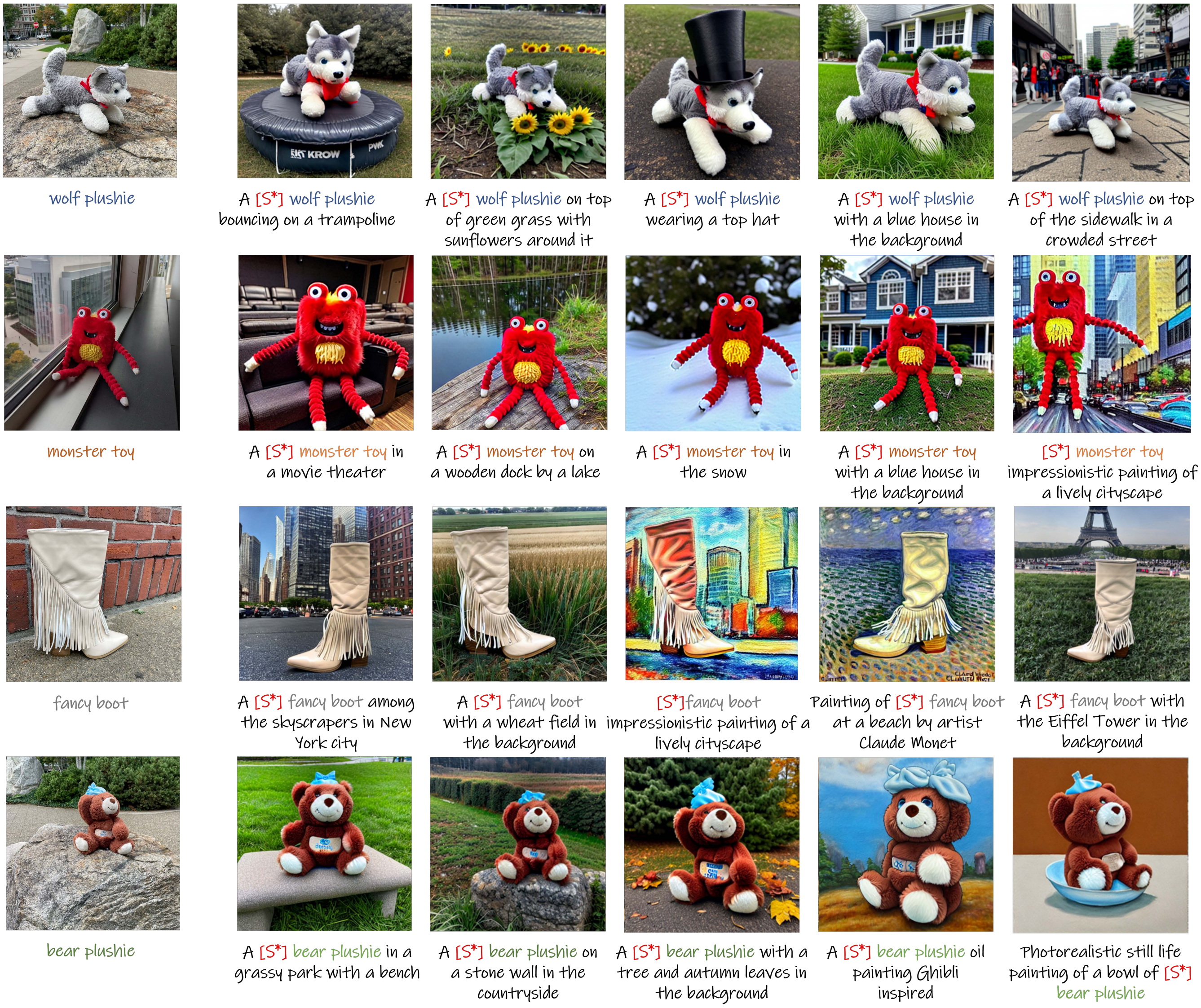

dog
A [S*] dog climbing a tree
A [S*] dog in a dream of a distant galaxy
A [S*] dog leaping across rooftops in a city
A [S*] dog wearing a black top hat and a monocle
A [S*] dog wearing pink glasses
dog
A [S*] dog diving into a pool
A [S*] dog on a cobblestone street
A [S*] dog wearing a rainbow scarf
A [S*] dog with a beautiful sunset
A [S*] dog with a ountain in the background
vase
A [S*] vase floating on top of water
A [S*] vase in a luxurious interior living room
Painting of [S*] vase at a beach by artist Claude Monet
Sketch of a [S*] vase
Watercolor painting of [S*] vase
duck toy
A [S*] duck toy at a picnic spot with a checkered blanket
A [S*] duck toy with the Eiffel Tower in the background
A purple [S*] duck toy
A mosaic representation of a [S*] duck toy
A surreal [S*] duck toy sculpture in bronze

Figure 6: More visual results of our method under different prompts.



wolf plushie

A [S*] wolf plushie bouncing on a trampoline

A [S*] wolf plushie on top of green grass with sunflowers around it

A [S*] wolf plushie wearing a top hat

A [S*] wolf plushie with a blue house in the background

A [S*] wolf plushie on top of the sidewalk in a crowded street

monster toy

A [S*] monster toy in a movie theater

A [S*] monster toy on a wooden dock by a lake

A [S*] monster toy in the snow

A [S*] monster toy with a blue house in the background

[S*] monster toy impressionistic painting of a lively cityscape

fancy boot

A [S*] fancy boot among the skyscrapers in New York city

A [S*] fancy boot with a wheat field in the background

[S*]fancy boot impressionistic painting of a lively cityscape

Painting of [S*] fancy boot at a beach by artist Claude Monet

A [S*] fancy boot with the Eiffel Tower in the background

bear plushie

A [S*] bear plushie in a grassy park with a bench

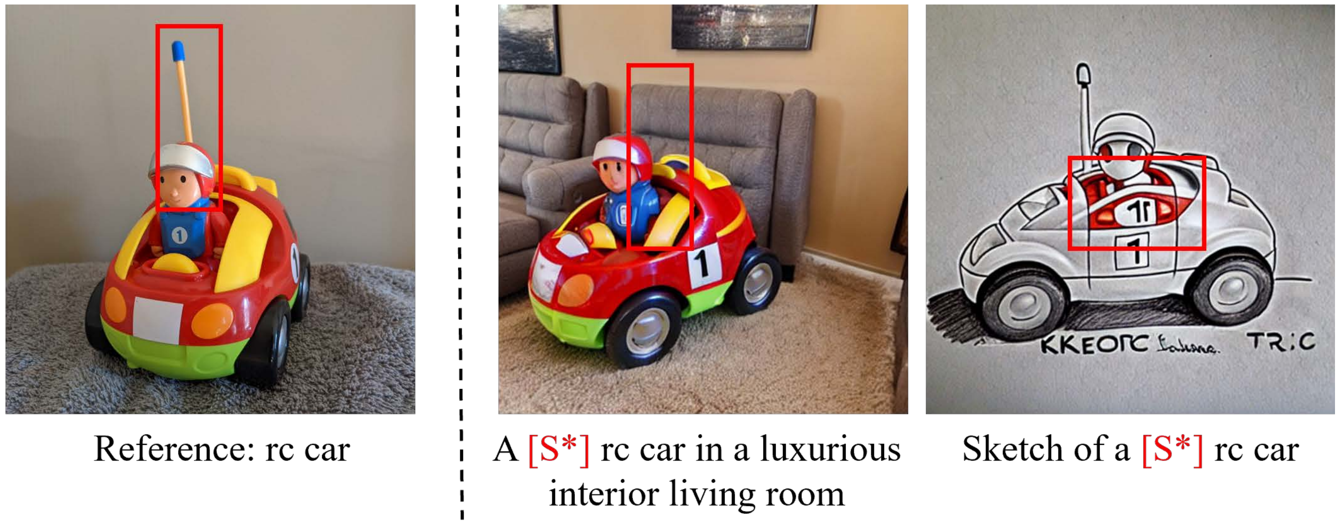
A [S*] bear plushie on a stone wall in the countryside

A [S*] bear plushie with a tree and autumn leaves in the background

A [S*] bear plushie oil painting Ghibli inspired

Photorealistic still life painting of a bowl of [S*] bear plushie

Figure 7: More visual results of our method under different prompts.



Reference: rc car

A [S*] rc car in a luxurious interior living room

Sketch of a [S*] rc car

Figure 8: Failure cases.