Selectively Informative Description can Reduce Undesired Embedding Entanglements in Text-to-Image Personalization

Supplementary Material

A. Evaluation measures

To evaluate embedding entanglements, we introduced two measures: *subject-alignment* and *non-subject-disentanglement*. These measures are obtained from the widely used image-alignment score, which calculates the cosine similarity between reference and generated images in the CLIP [13] (or DINO [1]) embedding space. Unlike the image-alignment score, subject-alignment and non-subject-disentanglement utilize segmented reference images to assess similarities associated with subject segments or non-subject segments, providing clear insights into both subject preservation and the presence of undesired entanglements.

In Fig. A.1, we compare image-alignment measure with our two customized measures using a dog example. The second and third rows in the figure display scores for two images generated by a personalized text-to-image diffusion model. The first row shows an image that fails to preserve the subject identity, with an undesirable background entanglement (natural landscape). The second row shows an image that successfully preserves the subject identity while disentangling the undesirable background. Our example in Fig. A.1 reveal that the image-alignment can be significantly influenced by background information, raising concerns about its accuracy when measuring subject preservation in highly biased scenarios. In contrast, subject-alignment accurately evaluates subject preservation and non-subject-disentanglement effectively identifies background entanglement in Fig. A.1. These differences make the two customized measures well-suited for analyzing text-to-image personalization models in any level of biased scenarios.

While the two customized measures effectively capture subject preservation and undesired entanglements, they have limitations in evaluating style re-contextualization. Specifically, the two measures rely on segmentation in the image pixel space, making it challenging to separate style from the image itself. Moreover, quantifying style similarities while disregarding the content presented in the image is not a straightforward task. Consequently, in all the quantitative analyses presented in this paper, we have excluded scenarios involving painting/cartoon style re-contextualization, even though our approach demonstrates significant visual improvements. As part of our future work, we plan to explore and identify a suitable measure capable of capturing style similarities independently of the image's content.

All three measures employed in this paper, namely subject-alignment, non-subject-disentanglement, and text-alignment, were calculated with the CLIP ViT-B/32 model. When calculating subject-alignment, we apply center alignment and image resizing to the subject segments before feeding them into the CLIP image encoder.

B. Human evaluation

Method	Subject-alignment	Non-subject-disentanglement	Text-alignment	Overall
DreamBooth	36.2%	10.9%	7.8%	15.5%
DreamBooth + SID	**40.6%**	**68.8%**	**59.7%**	**70.9%**
Undecided	23.2%	20.3%	32.5%	13.6%

Table B.1. Human evaluation results.

We conducted human evaluation involving 130 participants mainly recruited from our university community anonymously. The survey comprised 4 questions per subject, assessing subject-alignment, non-subject disentanglement, text-alignment, and the overall judgement considering all three aspects. Each participant evaluated 10 subjects, resulting in 40 problems, choosing between two images generated by DreamBooth and SID-integrated DreamBooth, respectively, and the order of the images was randomized. In cases of difficulty, participants could opt for the “Cannot Determine / Both Equally” option. In Tab. B.1, the survey results show that our method was preferred over DreamBooth in all aspects, with particularly notable differences observed in *text-alignment*, *non-subject disentanglement*, and the overall judgement. The results were also consistent with the results of our metric analysis in Fig. 8, supporting our metric analysis results significantly.

C. Implementations and datasets in detail

C.1. Descriptions in style re-contextualization

In style re-contextualization, we employed the phrase “A painting/cartoon in the style of [v] art’’ as our baseline train description, which is also used in Textual Inversion [2] and Custom Diffusion [6] to capture the styles of reference images. For generating SID that extends the baseline description, we made a slight modification to the VLM instruction as follows:

```
Describe the image in one sentence
in detail. Please start the sentence
with "A painting/cartoon in the style
of art.". You should not describe the
style of the painting/cartoon itself.
```

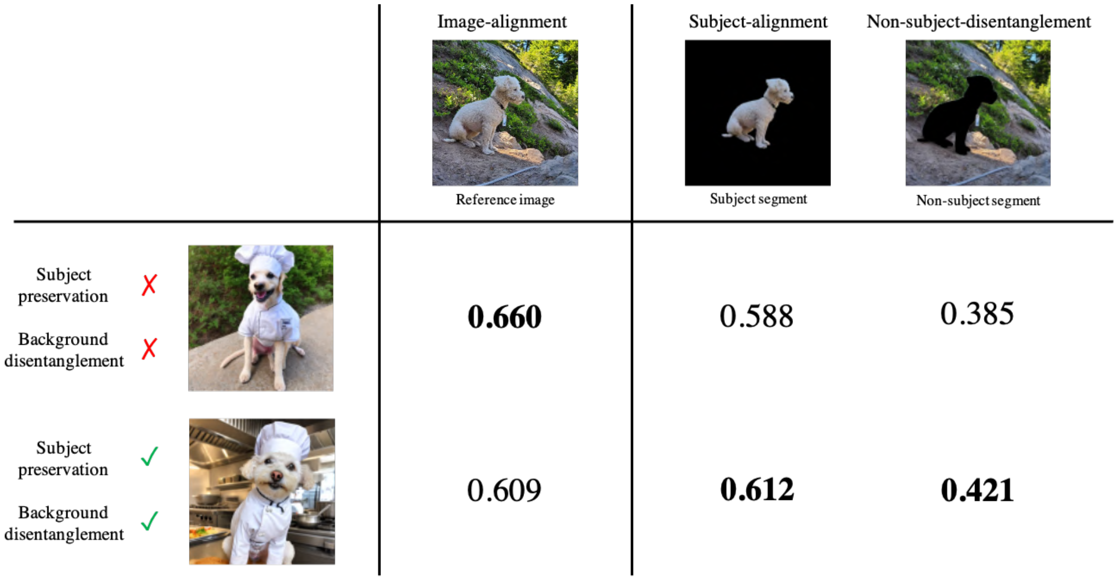


Figure A.1. **A comparison between image-alignment and two customized measures.** The top row shows the reference image and its two segmented versions, while the left column shows two images generated with the prompt “A [v] dog in a chef outfit.” Despite poor subject preservation, the upper image in the left column obtains a higher image-alignment score because of the background bias. In contrast, subject-alignment effectively assesses subject preservation and non-subject-disentanglement accurately identifies background disentanglement.

C.2. Model implementations

In this section, we provide the implementation details of the baseline models employed in this paper. We selected “stable-diffusion-2-1-base” as the backbone models for DreamBooth [15], SVDiff [3], Custom Diffusion [6], and Textual Inversion [2]. As for ELITE [20] and BLIP-Diffusion [7], we employed the backbone models used in their respective official GitHub repositories [16, 21]. All the generated images are sampled using the DDIM [18] sampler, employing 50 steps and a guidance scale [5] of 7.5.

DreamBooth [15]. We implemented DreamBooth using the Hugging Face Diffusers library [19]. The model was trained with a batch size of 1 and a learning rate of 1×10^{-6} for 1000 iterations. We also trained a CLIP [13] text encoder in conjunction with the U-Net [14].

SVDiff [3]. We implemented SVDiff using a third-party implementation [10] due to the unavailability of official code. The model was trained for 1000 iterations with default hyperparameter settings.

Custom Diffusion [6] & Textual Inversion [2]. We implemented Custom Diffusion and Textual Inversion using the Hugging Face Diffusers library [19], with default hyperparameter settings.

ELITE [20] & BLIP-Diffusion [7]. We implemented ELITE and BLIP-Diffusion using the official GitHub repositories [16, 21].

C.3. Dataset details

The quantitative analyses in our work are based on the dataset shown in Fig. C.1. The 15 subjects in the dataset were chosen as the prominently utilized subjects in three recent text-to-image personalization works [2, 6, 15]. Specifically, 7 subjects were chosen from DreamBooth [15]: `[cat, dog3, dog6, pink sunglasses, vase, backpack, rc car]`, 7 subjects from Custom Diffusion [6]: `[barn, cat, dog, flower, chair, teddybear, wooden pot]`, and 1 from Textual Inversion [2]: `[cat toy]`. The subjects cover a variety of categories including living animals (5 subjects), plants (1 subject), buildings (1 subject), toys (3 subjects), artistic containers (2 subjects), and three other objects (3 subjects). For each of these subjects, 25 generation prompts are used, and 20 images are generated per prompt with different random noise initializations, resulting in a total of 7500 generated images.

C.4. VLM details

In Sec. 3.2., we selected the multimodal GPT-4 as our choice for generating SID after evaluating three instruction-following VLMs, each with the following versions:

- **BLIP-2 [8]**: “blip2-opt-2.7b” [17]
- **LLaVA [9]**: “llava-v1-0719-336px-lora-merge-vicuna-13b-v1.3” [4]
- **GPT-4 [12]**: Multi-modal GPT-4 provided in ChatGPT [11]

We have noticed that the GPT-4 Vision API was announced to be released after November 6th, 2023, and we look forward to using it to simplify the SID generation process.

Given the rapid evolution of VLMs, including the recent releases of GPT-4 Turbo with vision and LLaVA v1.5, we acknowledge that our analysis in Sec. 3.2. is limited to the versions mentioned above. Comparison results may be affected with the introduction of these new VLM versions or customized VLM instructions. We emphasize that our primary research focus is to identify a description format that can reduce undesired entanglement rather than the ease of generating train descriptions. We acknowledge that the optimal VLM for generating SID can change with the emergence of newer VLM versions at any moment.

D. Limitations of SID

D.1. Failure cases with GPT-4

While GPT-4 [12] generally generates descriptions that closely match the provided instructions, occasional failures do occur. In Fig. D.1, we report two instances where GPT-4 does not generate prompts that align closely with the properties of SID.

D.2. Selectively describing facial expression

We observed that SID-integrated models often face difficulties in altering the subject's facial expression according to the generation prompt, especially when the reference images display strong facial expressions. This challenge arises due to the lack of text descriptions of the subject's facial expressions in the VLM-generated SID, resulting in undesired entanglement of the facial expressions in the subject embedding. We discovered that this challenge can be addressed by incorporating text descriptions of the facial expressions into the SID. Specifically, this can be achieved by modifying the instructions of VLM for SID or manually adding descriptions of the facial expressions into SID. For example, in Fig. D.2, we demonstrated that manually adding descriptions of the subject's facial expression can resolve the undesired feature entanglement issue, successfully altering the subject's facial expression.

E. Additional experiment results

In this section, we provide additional experiment results to augment those presented in the main text. Additionally, we conduct qualitative comparisons between DreamBooth [15] and its SID-integrated counterpart in scenarios with varying levels of biases.

E.1. Four description cases

To further validate the effectiveness of Case 3 (Ours) among the four description cases defined in Tab. 1, we conducted a quantitative comparison. The results are reported in Tab. E.2 and additional qualitative comparisons are shown in Fig. E.1. In Tab. E.2, Case 1 (Baseline) exhibits high subject alignment but lacks in text-alignment and non-subject disentanglement, indicating undesired embedding entanglement. On the other hand, Case 4 demonstrates optimal text-alignment and high non-subject disentanglement but significantly falls short in subject-alignment, as also evident in Fig. E.1. In contrast, both Case 2 and 3 achieve decent scores across all three measures, with Case 3 (Ours) outperforming Case 2 in all aspects. To conclude, Case 3 (Ours) achieves the best subject-alignment and non-subject-disentanglement, along with near-best text-alignment scores, which agree with the additional qualitative comparisons in Fig. E.1.

E.2. Instruction-following VLMs

To support the explanations presented in Sec. 3.2., we present additional comparisons of the three instruction-following VLMs in Fig. E.3. Additionally, we provide quantitative comparisons among the three VLMs to assess whether the multi-modal GPT-4 [12] outperforms the other two VLMs, namely LLaVA [9] and BLIP-2 [8], in terms of text-alignment, subject-alignment, and non-subject-disentanglement. The results are shown in Tab. E.3, demonstrating that GPT-4 excel over the other two VLMs in all three measures.

E.3. Cross-attention map analysis

To extend the analysis presented in Sec. 5, we conducted the cross-attention map analysis on the four cases of train descriptions detailed in Fig. 3, top row. In Fig. E.2, our analysis focused on the nearby-object bias, which can similarly be applied to other biased scenarios. In Cases 1 and 2, the identifier [v] erroneously focuses on the nearby-object, 'the purse', in each generated image. Conversely, Cases 3 and 4 demonstrate a notable improvement in focusing on the subject, 'the perfume'. The primary distinction between these two groups of cases lies in the inclusion of informative specifications related to the non-subjects. This indicates that incorporating informative specifications can reduce the undesired focus or generation of the non-subjects. However, it is crucial to include informative specification *selectively*, as demonstrated in Case 4, where the inclusion of the informative specification related to the subject destroys its identity.

E.4. Enhancement by SID

We present comprehensive generation results of Fig. 5 in Fig. E.4 (a), (b), (c), and (d), where the generation outcomes are presented without any selection. Additional experimental results can be found in Fig. E.4 (e), (f), (g), and (h). We observed that, in the context of undesired embedding disentanglement, SVDiff stands out among other models. In case of Textual Inversion, the baseline model exhibits a low level of subject alignment, resulting in the SID-integrated model encountering a similar challenge.

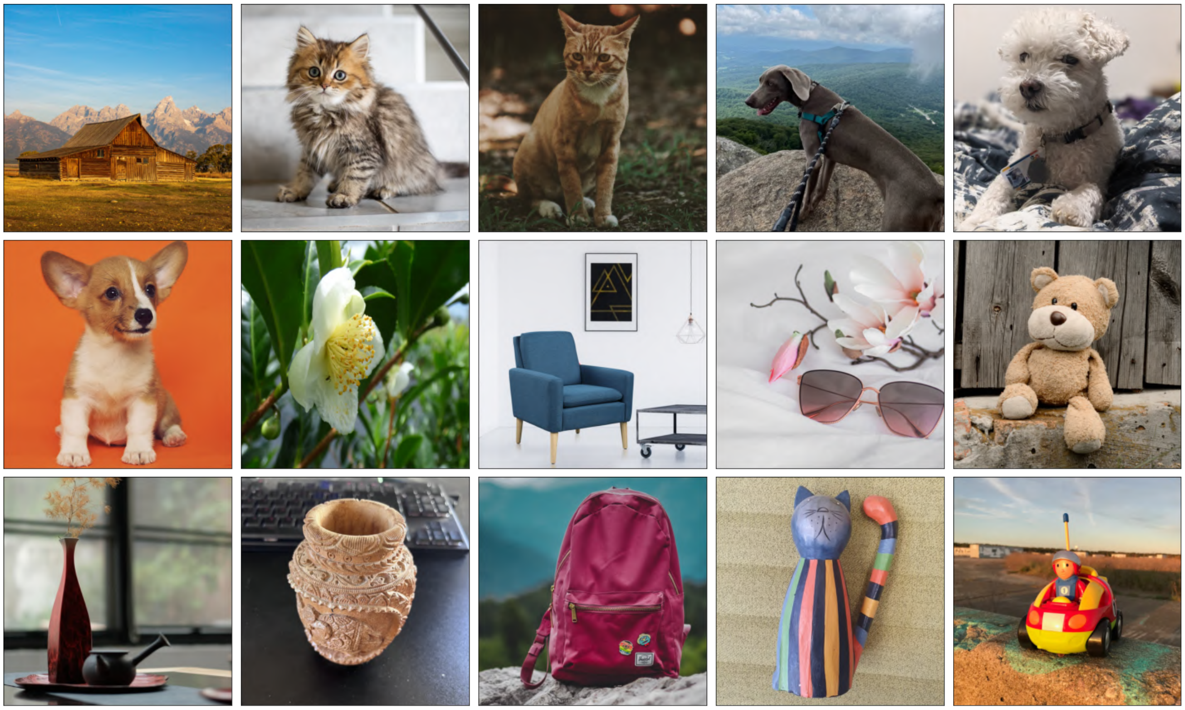


Figure C.1. **Dataset used for quantitative analyses.** An example image of each subject is shown. The 15 subjects were chosen from three recent works: 7 from DreamBooth [15], 7 from Custom Diffusion [6], and 1 from Textual Inversion [2].

Descriptions	Text alignment	Subject alignment	Non-subject disentanglement
Case 1 (Baseline)	0.290	**0.686**	0.299
Case 2	0.311	0.685	0.342
Case 3 (Ours)	0.317	**0.686**	**0.352**
Case 4	**0.325**	0.659	0.348

Table E.2. **Quantitative comparison of four description cases.** We have performed a quantitative analysis of the four cases of descriptions listed in Tab. 1. Case 3 (Ours) achieves the best performance for subject alignment and non-subject disentanglement. It also significantly improves text-alignment when compared to the Case 1 (Baseline).

E.5. Enhancement for a single reference image

We present comprehensive and additional results of Fig. 6 in Fig. E.5 (a), (b), (c), . . . , (h).

E.6. Negative prompt and segmentation

We present comprehensive and additional results of Fig. 10 in Fig. E.6 (a) and (b).

E.7. Highly, moderately, and low-biased scenarios

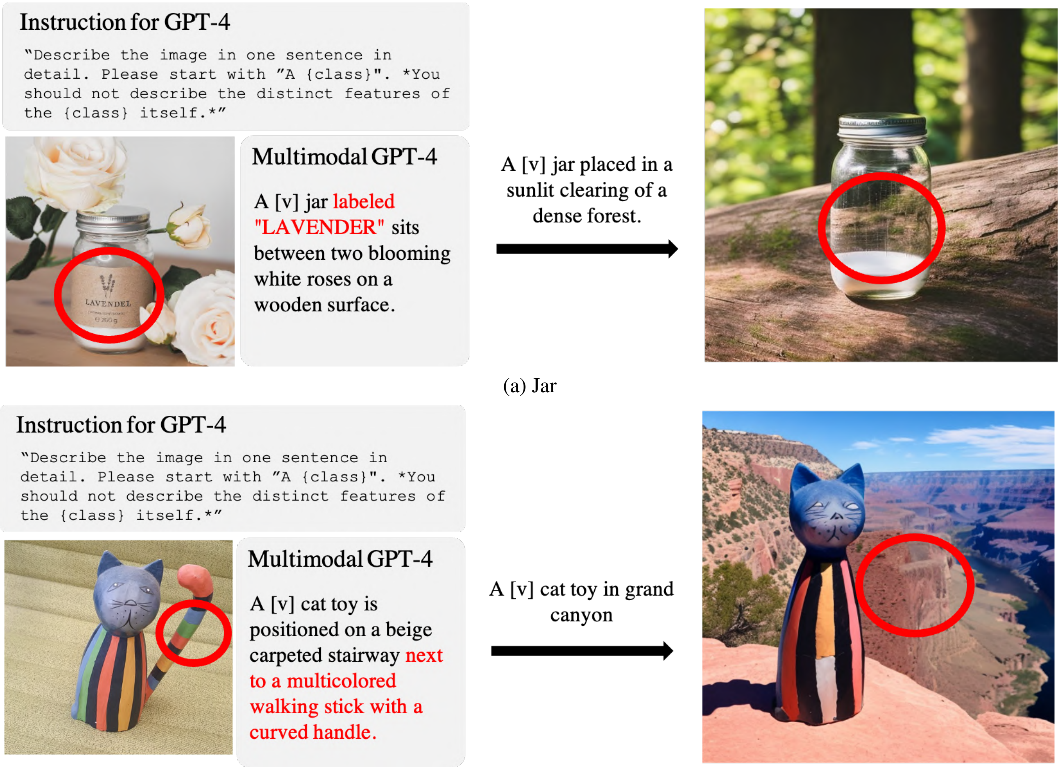
We compare DreamBooth [15] with its SID-integrated counterpart in scenarios with varying levels of biases: high, moderate, and low. The generation results for each scenario can be found in Fig. E.7 (a), (b), and (c).

VLMs	Text alignment	Subject alignment	Non-subject disentanglement
GPT-4 [12]	**0.317**	**0.684**	**0.354**
LLaVA [9]	0.311	0.681	0.340
BLIP-2 [8]	0.311	0.683	0.343

Table E.3. **Quantitative comparison of instruction-following VLMs.** The multi-modal GPT-4 demonstrates superior performance in all three measures when used for generating SIDs.

F. Societal impact

Our approach enhances personalized image synthesis, making it easier to create realistic images of personalized subjects in new contexts, even with highly biased reference images, including just a single reference image. While this advancement fosters creativity and contributes to the sharing of personalized content that closely aligns with user guidance, it also raises concerns about potential misuse by malicious users who may exploit generative models for deception or unauthorized copyright infringement. Additionally, these generative models inherit biases from the large-scale dataset used in the pre-training stage, which could inadvertently misinform the public. Future research should prioritize defending against such misuse and reducing biases in generative models to ensure responsible and ethical use, particularly in personalized image synthesis.



Instruction for GPT-4

"Describe the image in one sentence in detail. Please start with "A {class}". *You should not describe the distinct features of the {class} itself.*"

Multimodal GPT-4

A [v] jar labeled "LAVENDER" sits between two blooming white roses on a wooden surface.

A [v] jar placed in a sunlit clearing of a dense forest.

(a) Jar

Instruction for GPT-4

"Describe the image in one sentence in detail. Please start with "A {class}". *You should not describe the distinct features of the {class} itself.*"

Multimodal GPT-4

A [v] cat toy is positioned on a beige carpeted stairway next to a multicolored walking stick with a curved handle.

A [v] cat toy in grand canyon

(b) Cat toy

Figure D.1. **Imperfection of GPT-4.** In (a), the description generated by GPT-4 includes an informative specification of the label attached to the jar, leading to the omission of the subject's label in the generated image. In (b), the description depicts the tail of the cat toy as a walking stick, resulting in the omission of the tail in the generated image.

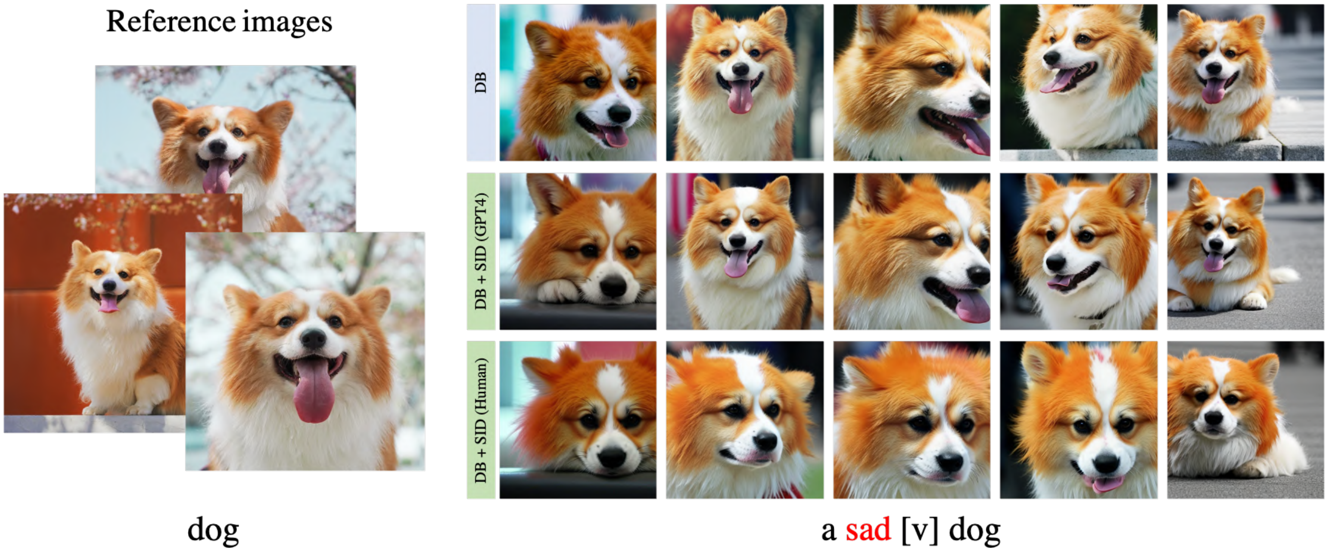


Figure D.2. **Selectively describing facial expression.** When the SID generated by GPT-4 lacks information about facial expression, the SID-integrated model fails to modify the subject's facial expression, as depicted in the middle row. By manually designing SID to describe the subject's facial expression, the SID-integrated model successfully disentangles the subject identity from the facial expression, as demonstrated in the last row. The manually designed SID is "a [v] dog with a joyful and open-mouthed expression, with its tongue hanging out."

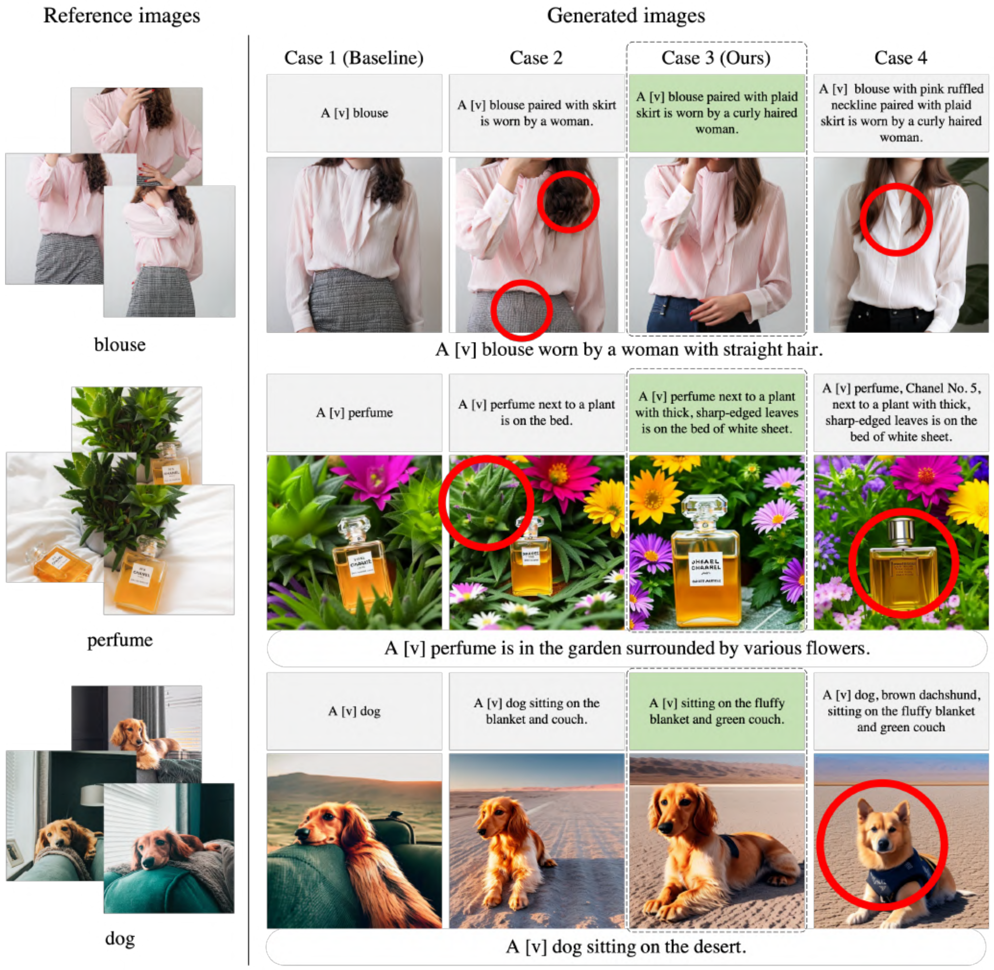

Reference images
Generated images
Case 1 (Baseline)
Case 2
Case 3 (Ours)
Case 4
A [v] blouse
A [v] blouse paired with skirt is worn by a woman.
A [v] blouse paired with plaid skirt is worn by a curly haired woman.
A [v] blouse with pink ruffled neckline paired with plaid skirt is worn by a curly haired woman.
blouse
A [v] blouse worn by a woman with straight hair.
A [v] perfume
A [v] perfume next to a plant is on the bed.
A [v] perfume next to a plant with thick, sharp-edged leaves is on the bed of white sheet.
A [v] perfume, Chanel No. 5, next to a plant with thick, sharp-edged leaves is on the bed of white sheet.
perfume
A [v] perfume is in the garden surrounded by various flowers.
A [v] dog
A [v] dog sitting on the blanket and couch.
A [v] sitting on the fluffy blanket and green couch.
A [v] dog, brown dachshund, sitting on the fluffy blanket and green couch
dog
A [v] dog sitting on the desert.

Figure E.1. **Additional examples for comparing the four cases of descriptions.** Case 2 shows a decent generation with occasional entanglements such as (top row) curly hair, not aligned with generation prompt, and gray plaid pattern and (middle row) spiky plant. Case 4 demonstrates high text-alignment but significantly falls short in subject preservation. Case 3 (Ours) successfully achieves the desired qualities: text-alignment, subject-alignment, and non-subject-disentanglement.

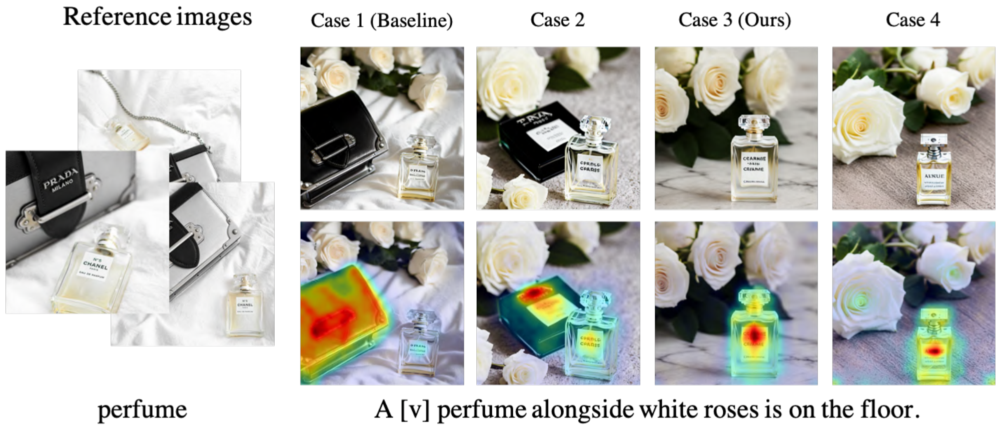
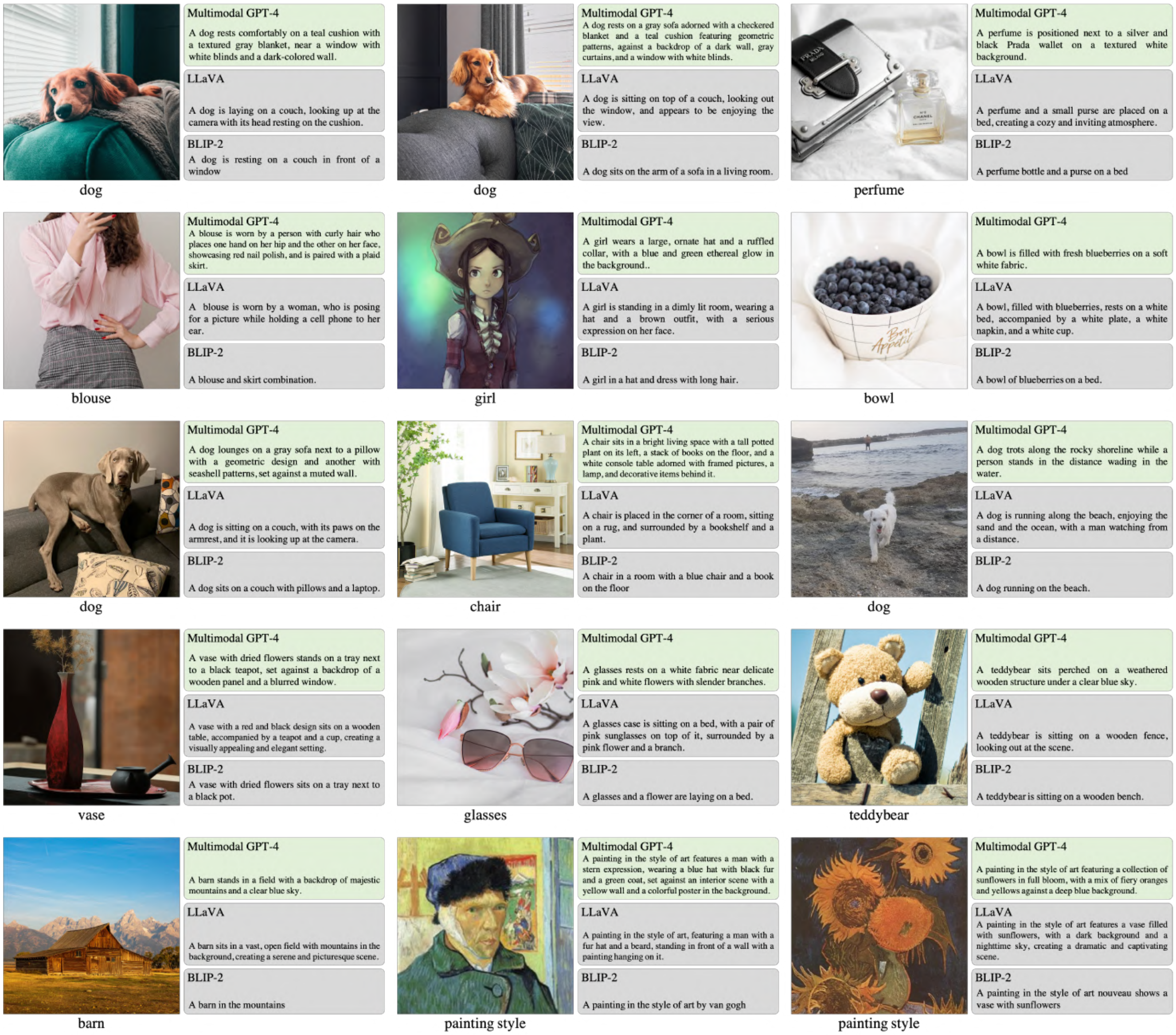


Figure E.2. **Cross-attention maps of the identifier [v] for the four cases of train descriptions from Fig. 3, top row.** In Cases 1 and 2, attention is spread towards non-subject parts. However, in Cases 3 and 4, thanks to the informative specification of the non-subject, this spreading of attention is highly reduced.



Multimodal GPT-4
A dog rests comfortably on a teal cushion with a textured gray blanket, near a window with white blinds and a dark-colored wall.
LLaVA
A dog is laying on a couch, looking up at the camera with its head resting on the cushion.
BLIP-2
A dog is resting on a couch in front of a window
dog

Multimodal GPT-4
A dog rests on a gray sofa adorned with a checkered blanket and a teal cushion featuring geometric patterns, against a backdrop of a dark wall, gray curtains,and a window with white blinds.
LLaVA
A dog is sitting on top of a couch, looking out the window, and appears to be enjoying the view.
BLIP-2
A dog sits on the arm of a sofa in a living room.
dog

Multimodal GPT-4
A perfume is positioned next to a silver and black Prada wallet on a textured white background.
LLaVA
A perfume and a small purse are placed on a bed, creating a cozy and inviting atmosphere.
BLIP-2
A perfume bottle and a purse on a bed
perfume

Multimodal GPT-4
A blouse is worn by a person with curly hair who places one hand on her hip and the other on her face, showcasing red nail polish, and is paired with a plaid skirt.
LLaVA
A blouse is worn by a woman, who is posing for a picture while holding a cell phone to her ear.
BLIP-2
A blouse and skirt combination.
blouse

Multimodal GPT-4
A girl wears a large, ornate hat and a ruffled collar, with a blue and green ethereal glow in the background..
LLaVA
A girl is standing in a dimly lit room, wearing a hat and a brown outfit, with a serious expression on her face.
BLIP-2
A girl in a hat and dress with long hair.
girl

Multimodal GPT-4
A bowl is filled with fresh blueberries on a soft white fabric.
LLaVA
A bowl, filled with blueberries, rests on a white bed, accompanied by a white plate, a white napkin, and a white cup.
BLIP-2
A bowl of blueberries on a bed.
bowl

Multimodal GPT-4
A dog lounges on a gray sofa next to a pillow with a geometric design and another with seashell patterns, set against a muted wall.
LLaVA
A dog is sitting on a couch, with its paws on the armrest, and it is looking up at the camera.
BLIP-2
A dog sits on a couch with pillows and a laptop.
dog

Multimodal GPT-4
A chair sits in a bright living space with a tall potted plant on its left, a stack of books on the floor, and a white console table adorned with framed pictures, a lamp, and decorative items behind it.
LLaVA
A chair is placed in the corner of a room, sitting on a rug, and surrounded by a bookshelf and a plant.
BLIP-2
A chair in a room with a blue chair and a book on the floor
chair

Multimodal GPT-4
A dog trots along the rocky shoreline while a person stands in the distance wading in the water.
LLaVA
A dog is running along the beach, enjoying the sand and the ocean, with a man watching from a distance.
BLIP-2
A dog running on the beach.
dog

Multimodal GPT-4
A vase with dried flowers stands on a tray next to a black teapot, set against a backdrop of a wooden panel and a blurred window.
LLaVA
A vase with a red and black design sits on a wooden table, accompanied by a teapot and a cup, creating a visually appealing and elegant setting.
BLIP-2
A vase with dried flowers sits on a tray next to a black pot.
vase

Multimodal GPT-4
A glasses rests on a white fabric near delicate pink and white flowers with slender branches.
LLaVA
A glasses case is sitting on a bed, with a pair of pink sunglasses on top of it, surrounded by a pink flower and a branch.
BLIP-2
A glasses and a flower are laying on a bed.
glasses

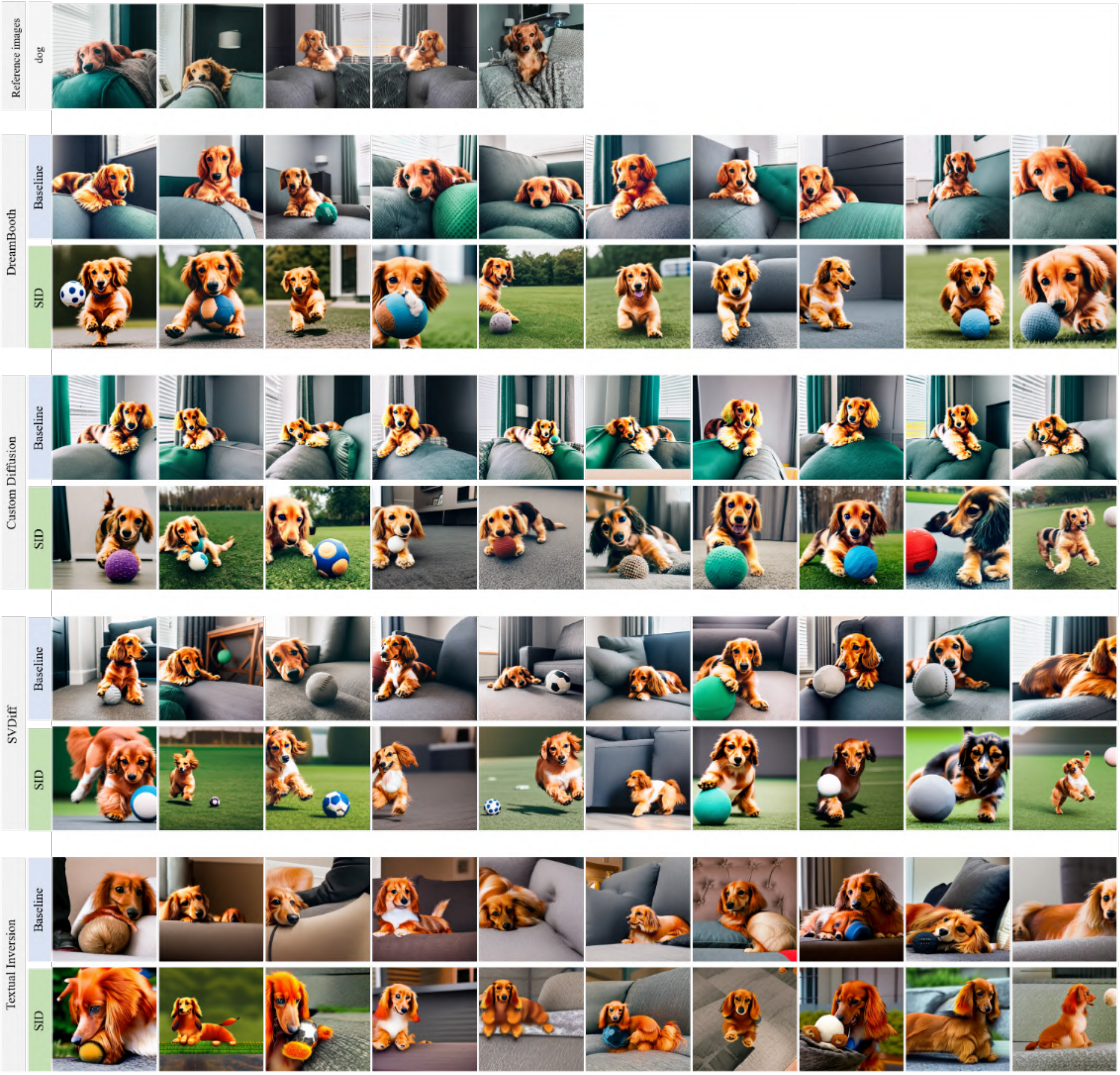
Multimodal GPT-4
A teddybear sits perched on a weathered wooden structure under a clear blue sky.
LLaVA
A teddybear is sitting on a wooden fence, looking out at the scene.
BLIP-2
A teddybear is sitting on a wooden bench.
teddybear

Multimodal GPT-4
A barn stands in a field with a backdrop of majestic mountains and a clear blue sky.
LLaVA
A barn sits in a vast, open field with mountains in the background, creating a serene and picturesque scene.
BLIP-2
A barn in the mountains
barn

Multimodal GPT-4
A painting in the style of art features a man with a stern expression, wearing a blue hat with black fur and a green coat, set against an interior scene with a yellow wall and a colorful poster in the background.
LLaVA
A painting in the style of art, featuring a man with a fur hat and a beard, standing in front of a wall with a painting hanging on it.
BLIP-2
A painting in the style of art by van gogh
painting style

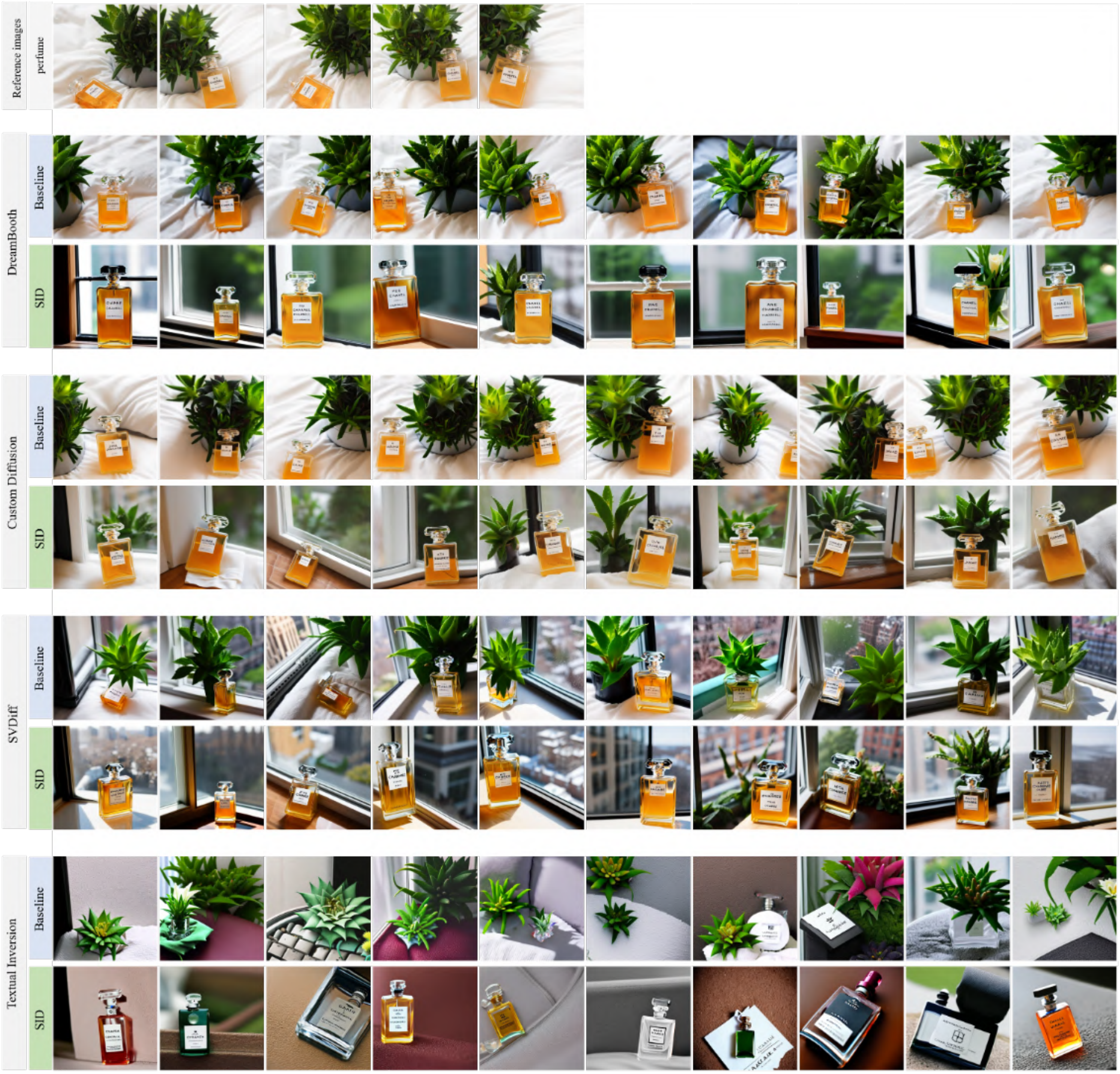
Multimodal GPT-4
A painting in the style of art featuring a collection of sunflowers in full bloom, with a mix of fiery oranges and yellows against a deep blue background.
LLaVA
A painting in the style of art features a vase filled with sunflowers, with a dark background and a nighttime sky, creating a dramatic and captivating scene.
BLIP-2
A painting in the style of art nouveau shows a vase with sunflowers
painting style

Figure E.3. **Additional examples for comparing the three instruction-following VLMs for generating SID.** LLaVA tends to include informative specifications of the subject itself or fall short in providing sufficient specifications of the undesired objects. BLIP-2 exhibits limitations in achieving a thorough identification of the undesired objects. Even when it successfully identifies an undesired object, it tends to generate simple captions without informative specifications. Compared to the other two, GPT-4 excels in generating captions that satisfy our instructions.



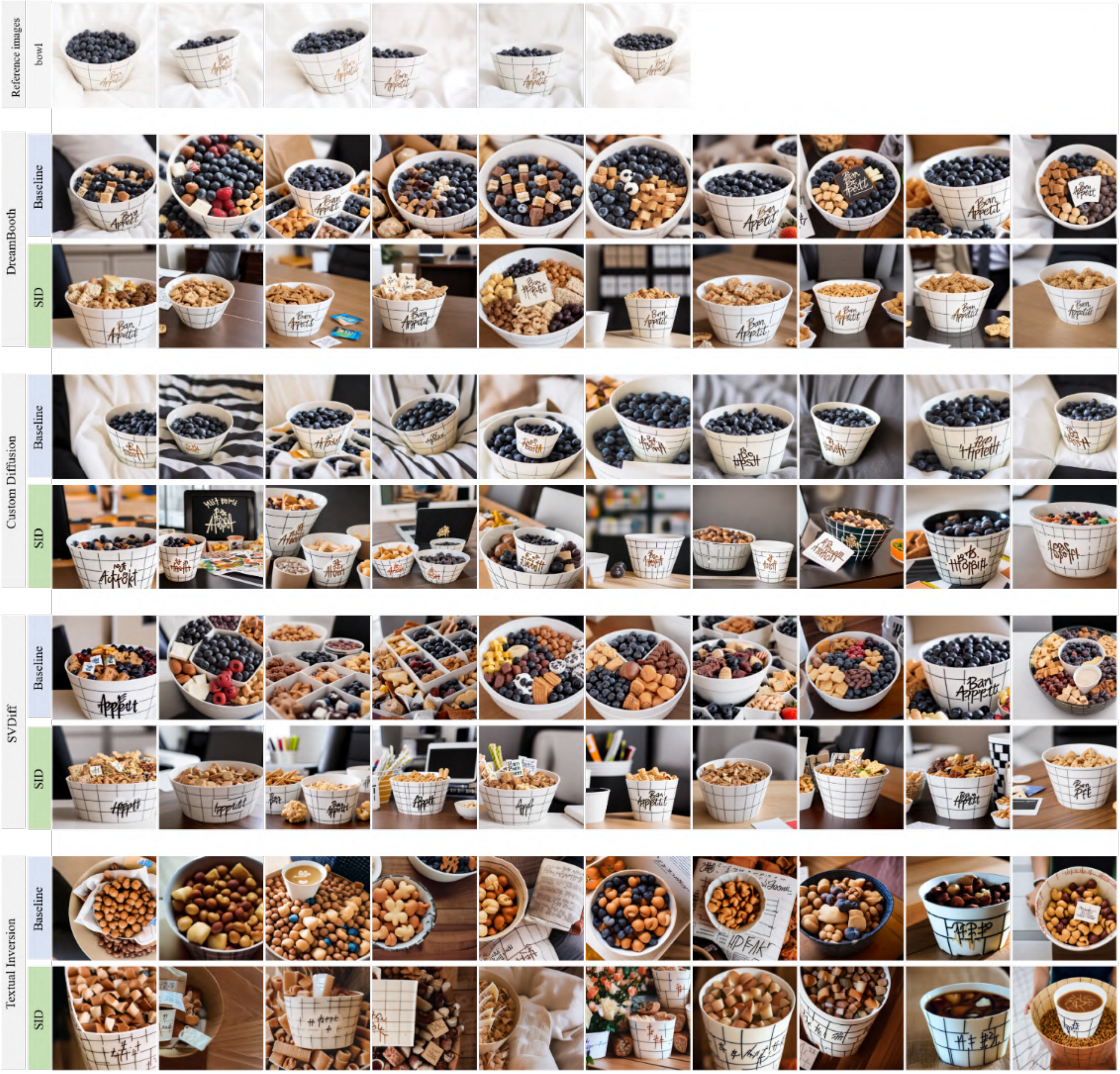
A [v] dog playing with a ball.

Figure E.4. (a) **Enhancement by SID – additional example #1.** SID-integration effectively resolves the entanglement problem of indoor background (background bias).



A [v] perfume on the windowsill.

Figure E.4. (b) **Enhancement by SID – additional example #2.** SID-integration effectively resolves the entanglement problems of potted plant and white sheet background (nearby-object bias).



A [v] bowl full of snacks is on the office table.

Figure E.4. (c) **Enhancement by SID – additional example #3.** SID-integration effectively resolves the entanglement problem of filled-in blueberries (tied-object bias).

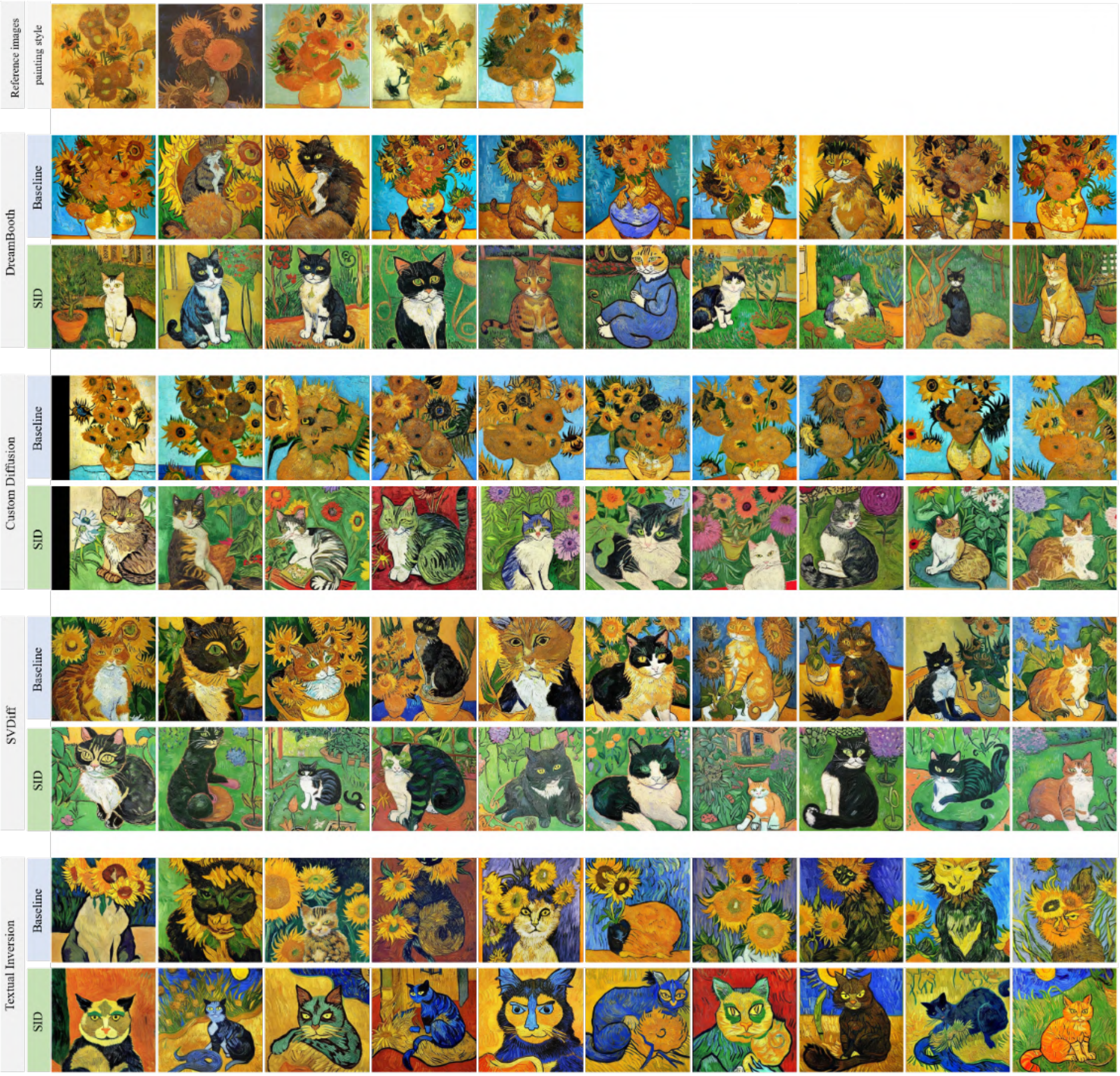


Figure E.4. (d) **Enhancement by SID – additional example #4.** SID-integration effectively resolves the entanglement problem of sunflowers (substance bias).

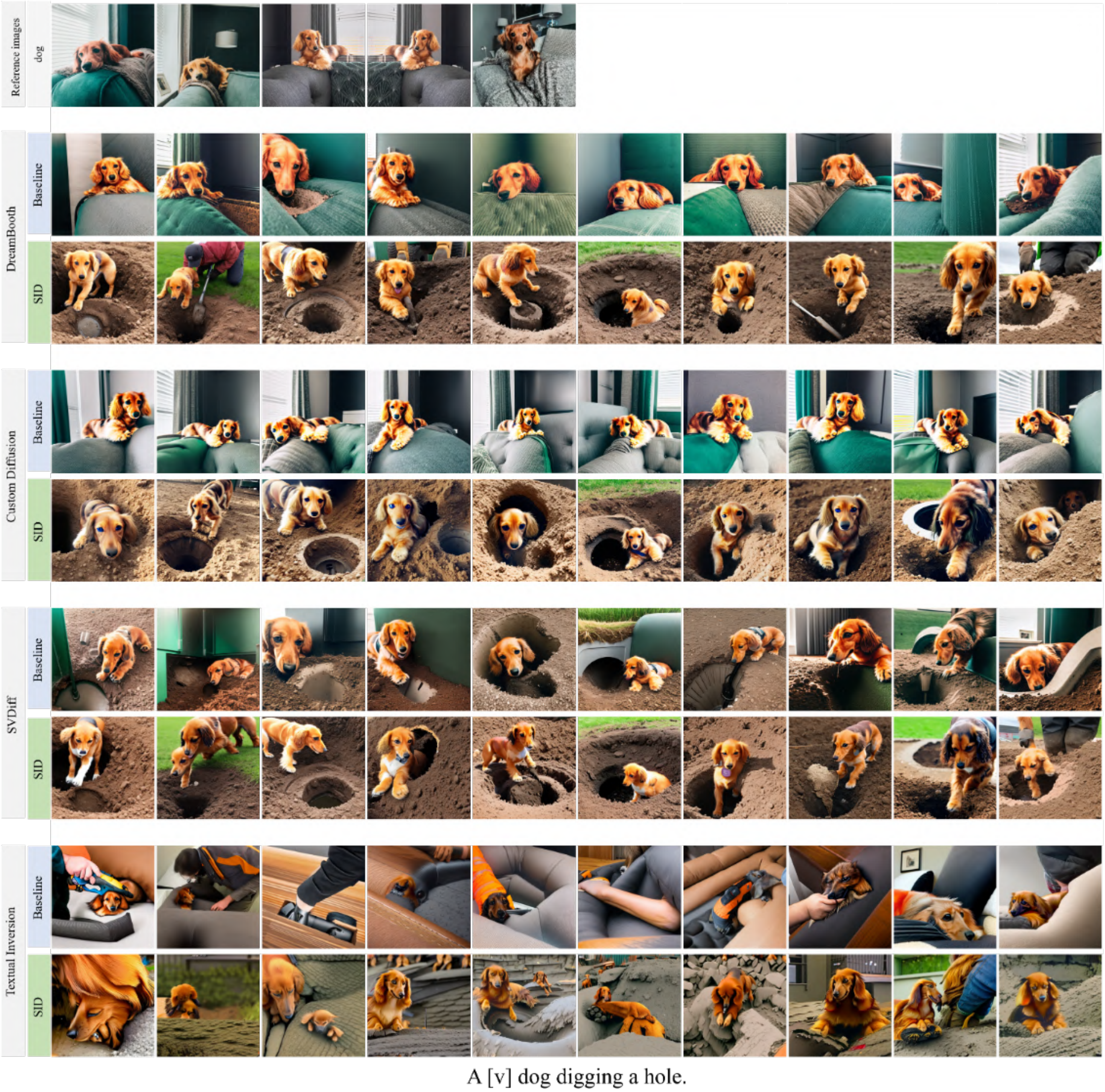
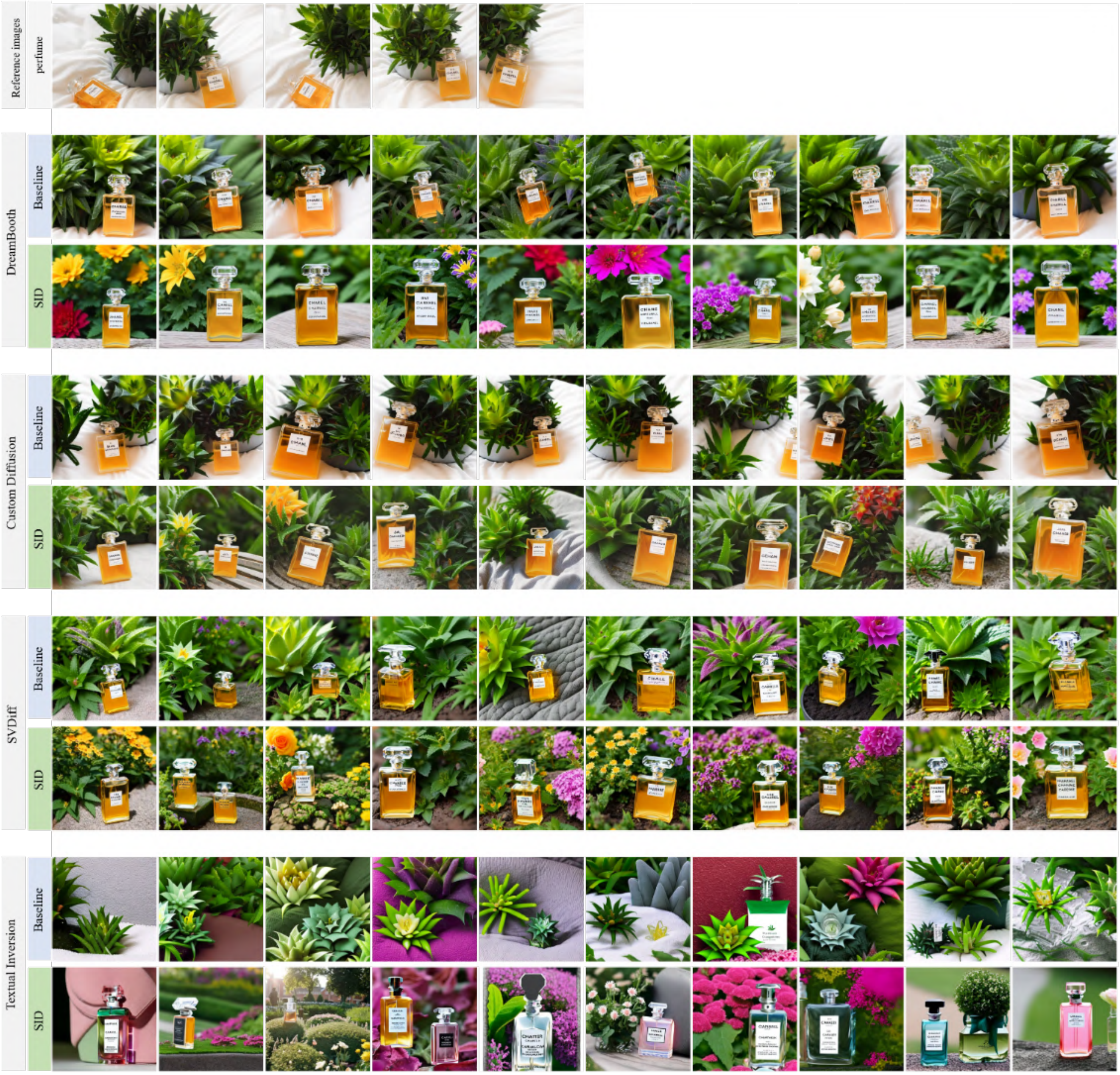


Figure E.4. (e) **Enhancement by SID – additional example #5.** SID-integration effectively resolves the entanglement problem of indoor background (background bias).



A [v] perfume in the garden.

Figure E.4. (f) **Enhancement by SID – additional example #6.** SID-integration effectively resolves the entanglement problem of a plant with sharp-edged leaves (nearby-object bias).

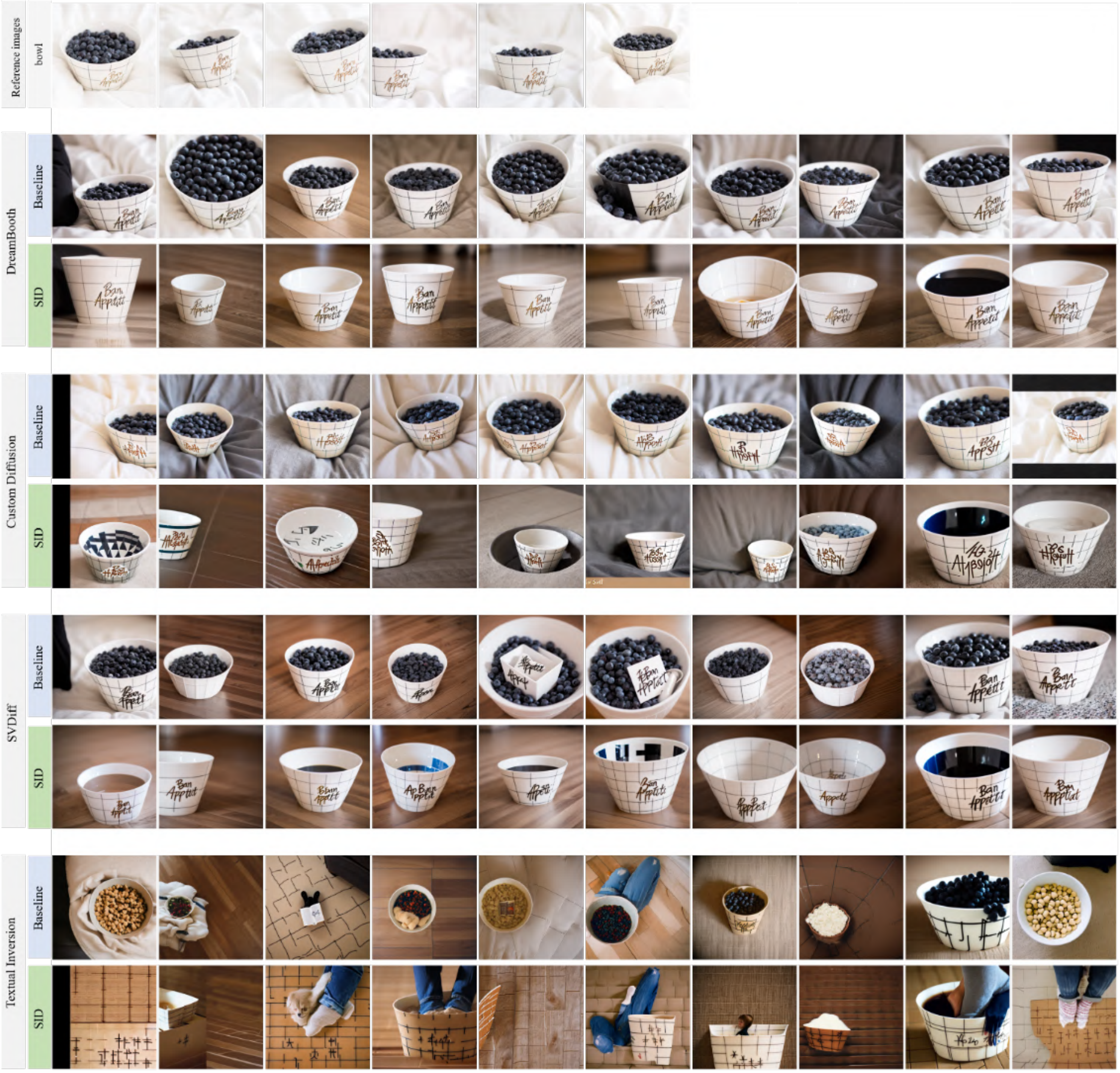


Figure E.4. (g) **Enhancement by SID – additional example #7.** SID-integration effectively resolves the entanglement problem of filled-in blueberries and a white sheet background (tied-object bias and background bias).

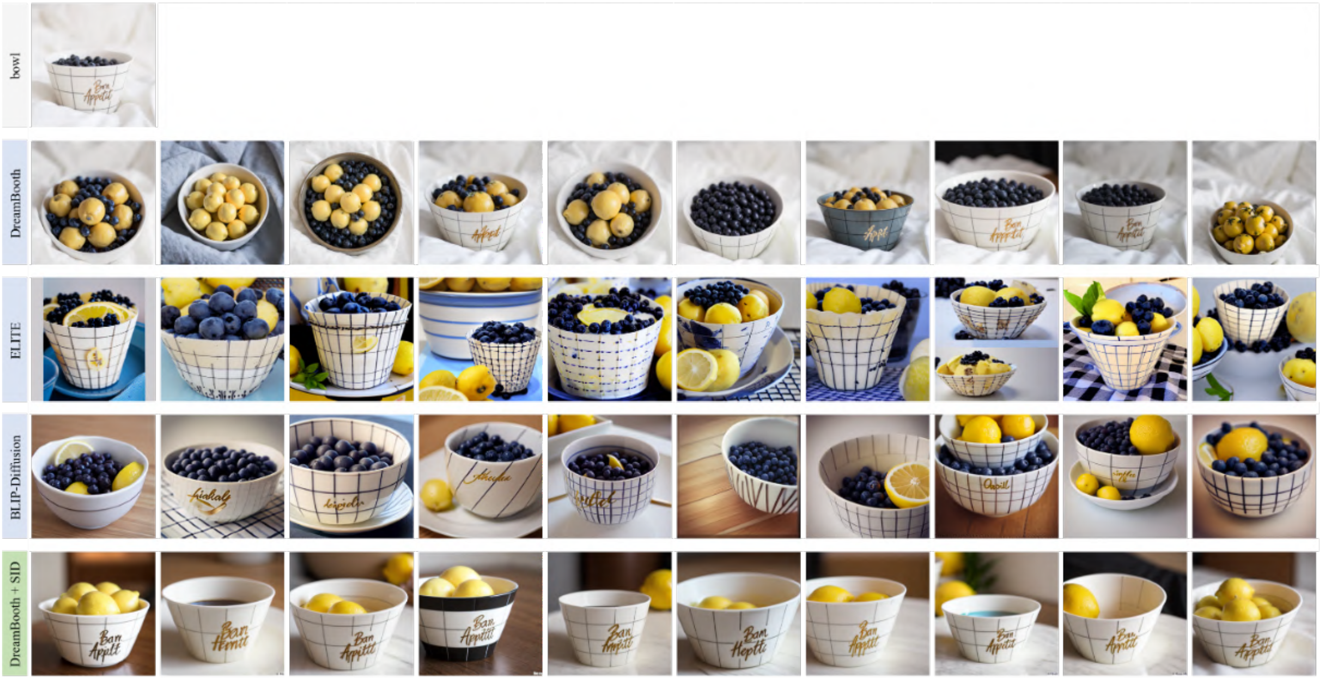


Figure E.5. (a) **Enhancement by SID for a single reference image – additional example #1.** DreamBooth and encoder-based models fail at removing blueberries from the bowl. Furthermore, they often fail at filling the bowl with lemons.

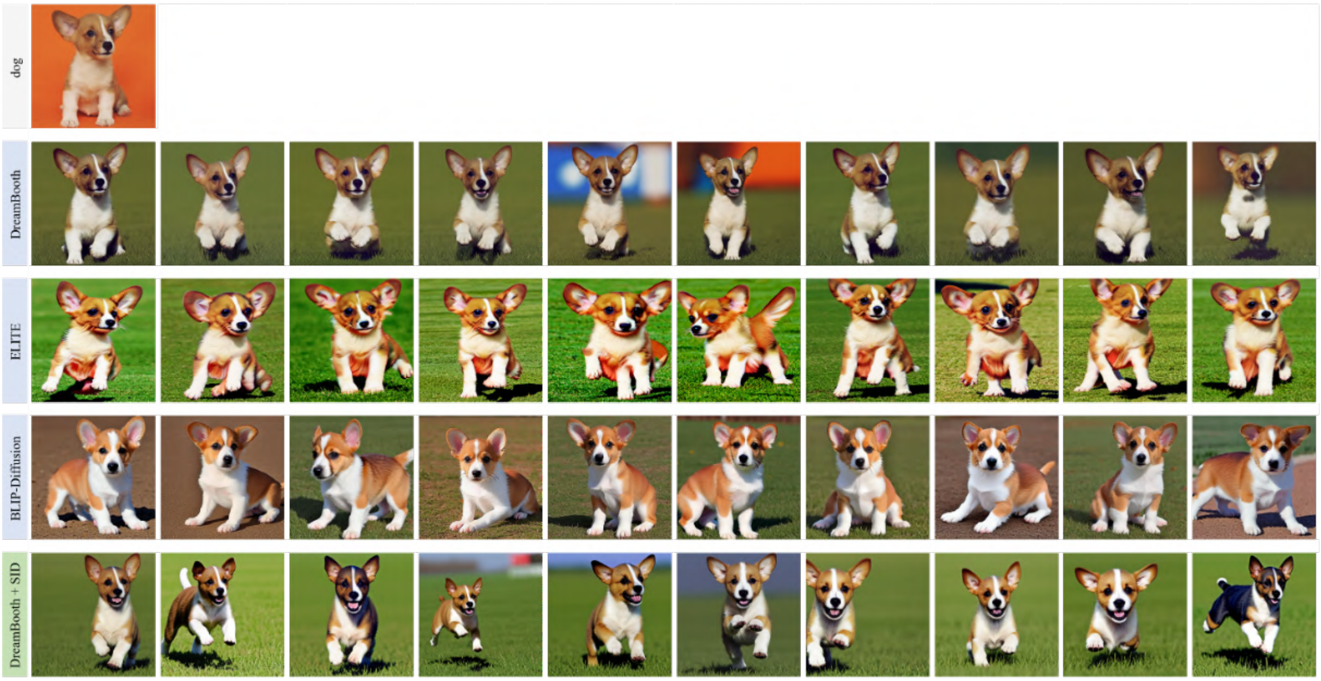


Figure E.5. (b) **Enhancement by SID for a single reference image – additional example #2.** DreamBooth and encoder based models struggle with generating images of a dog running in diverse poses (pose bias).

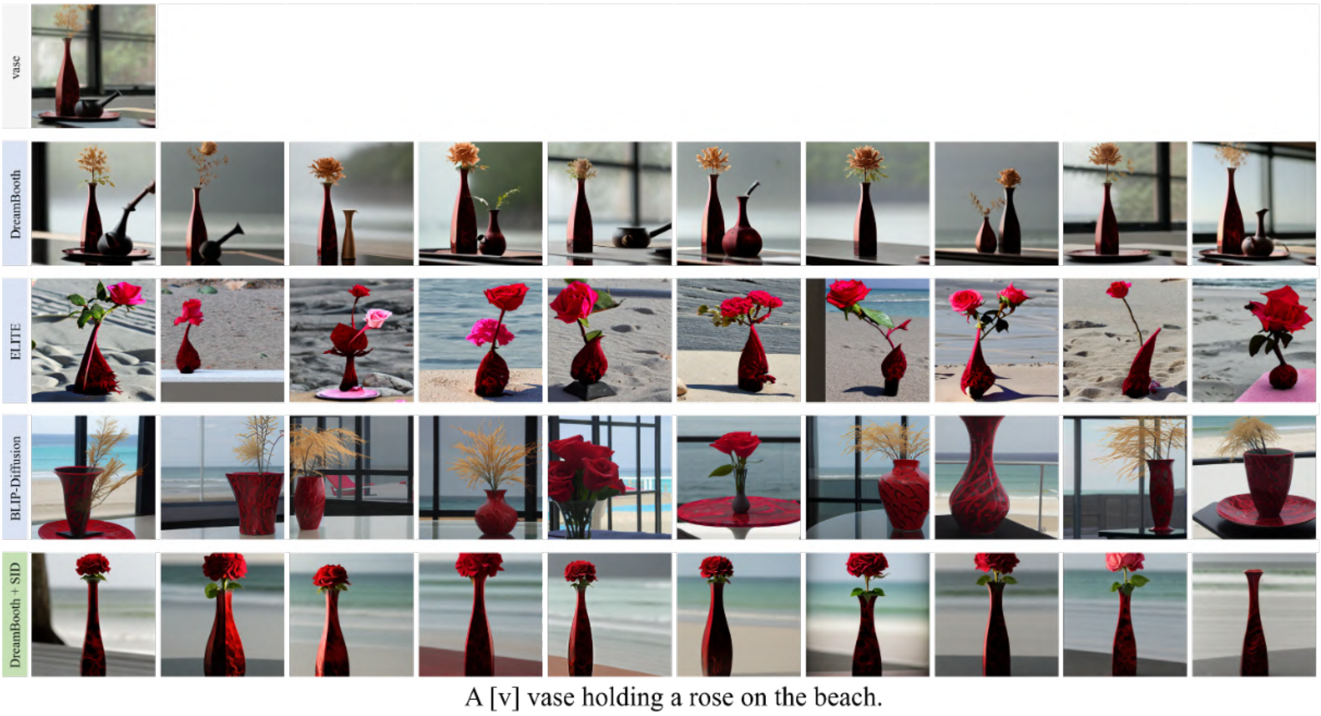


Figure E.5. (c) **Enhancement by SID for a single reference image – additional example #3.** DreamBooth not only generates a vase but also the nearby teapot and saucer. Encoder-based models struggle to preserve the identity of the vase, possibly because of the infrequent occurrences of vases during the encoder pre-training.

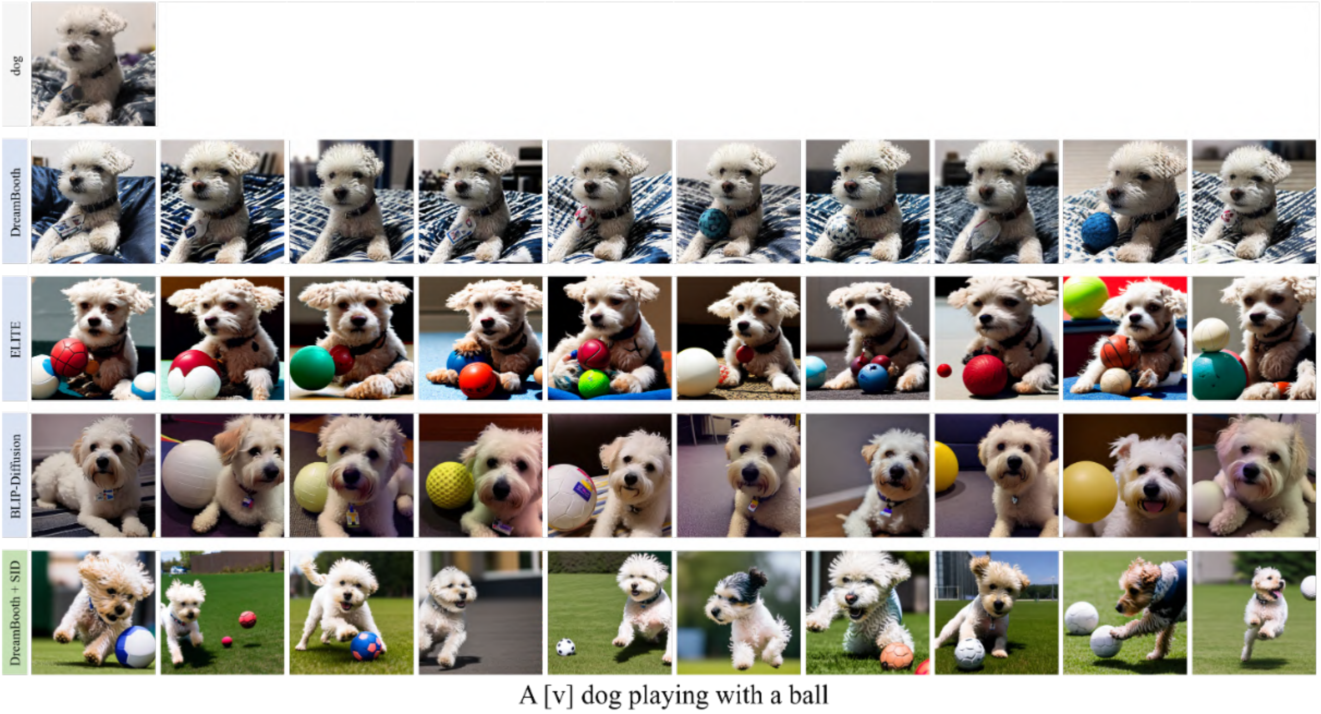


Figure E.5. (d) **Enhancement by SID for a single reference image – additional example #4.** DreamBooth and encoder based models struggle with generating images of a dog playing with a ball in diverse poses (pose bias).

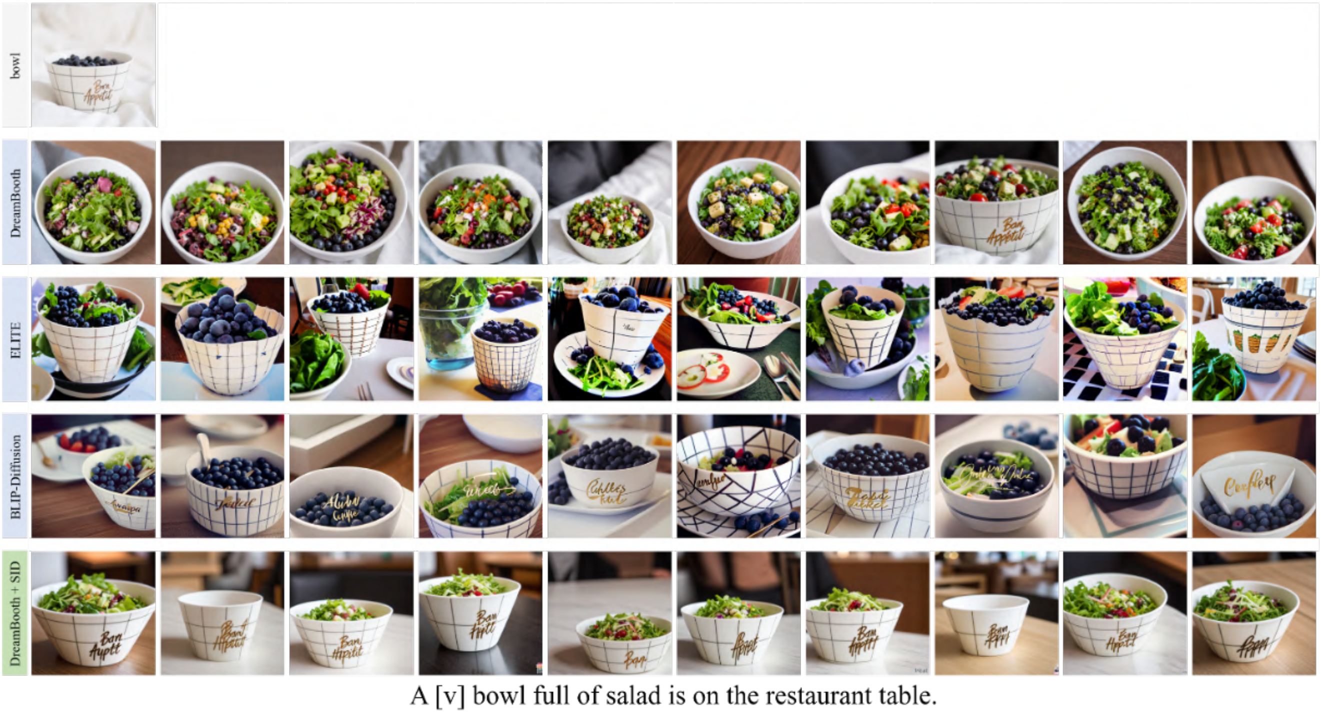


Figure E.5. (e) **Enhancement by SID for a single reference image – additional example #5.** DreamBooth struggles with preserving the identity of the bowl and also with generating a background that is aligned with the generation prompt. Encoder-based models face challenges in removing blueberries from the bowl and replacing them with salad.

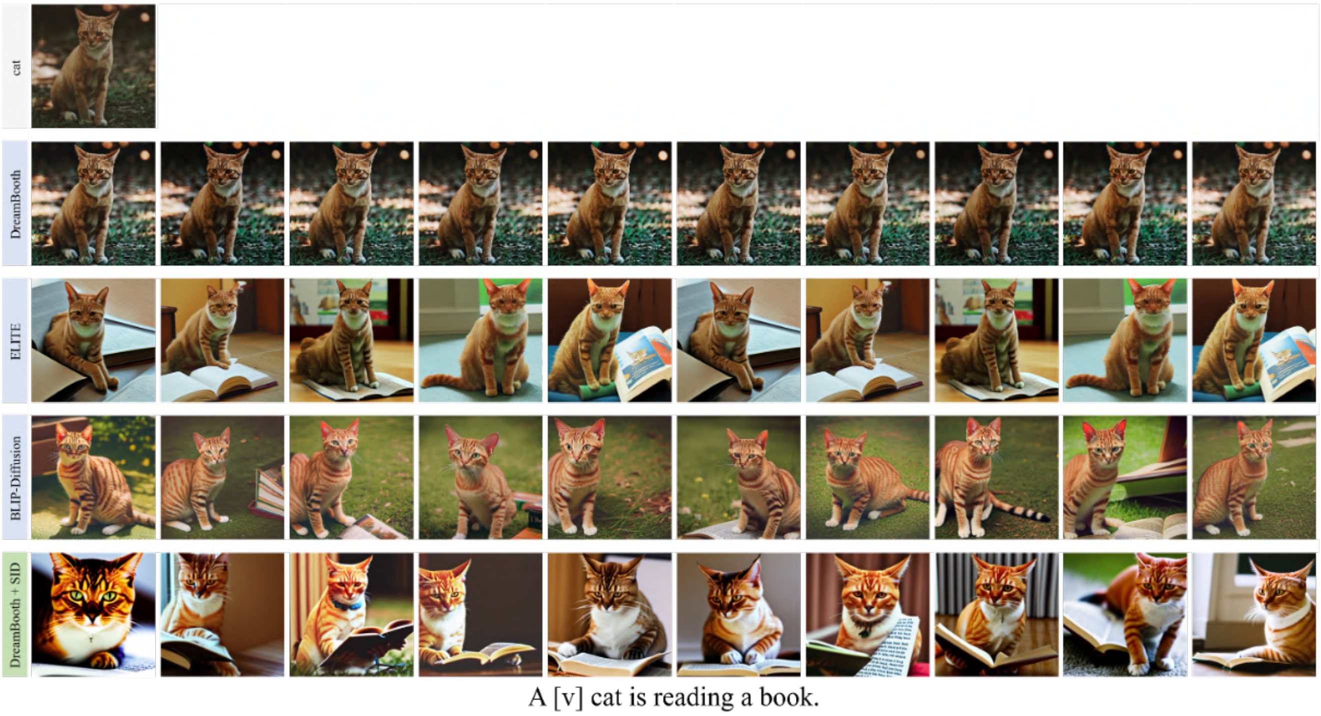
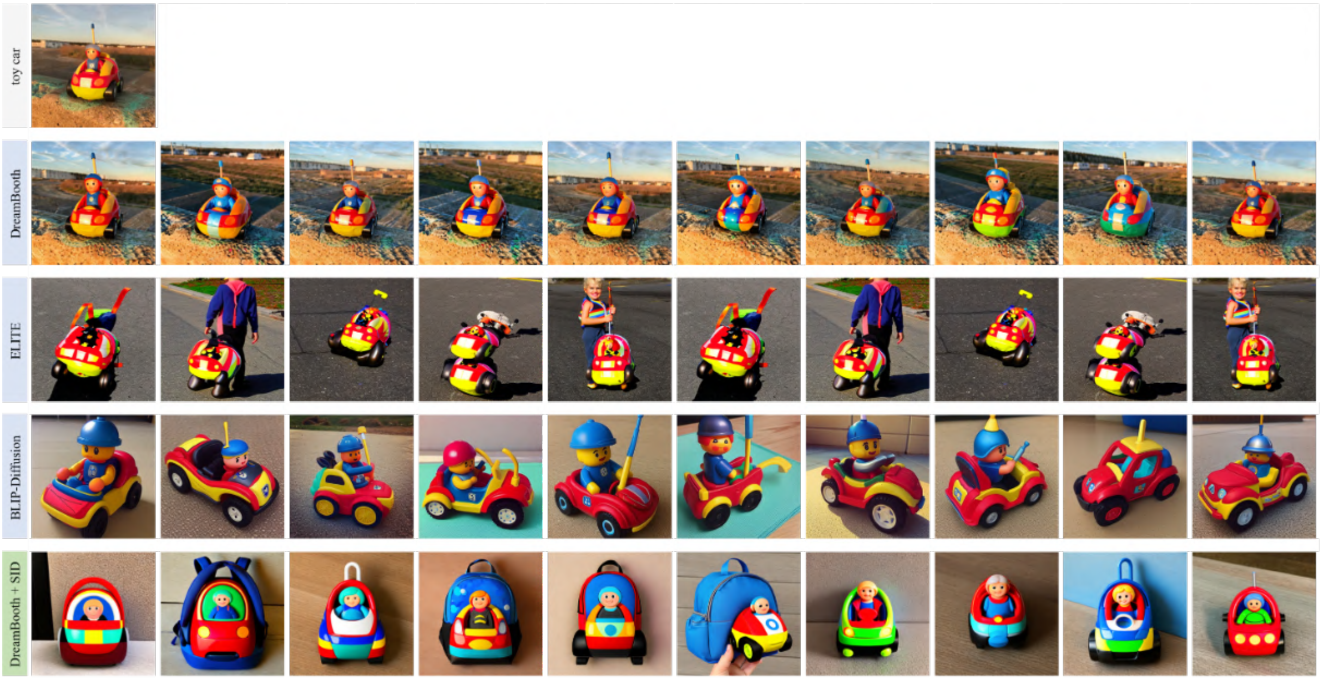
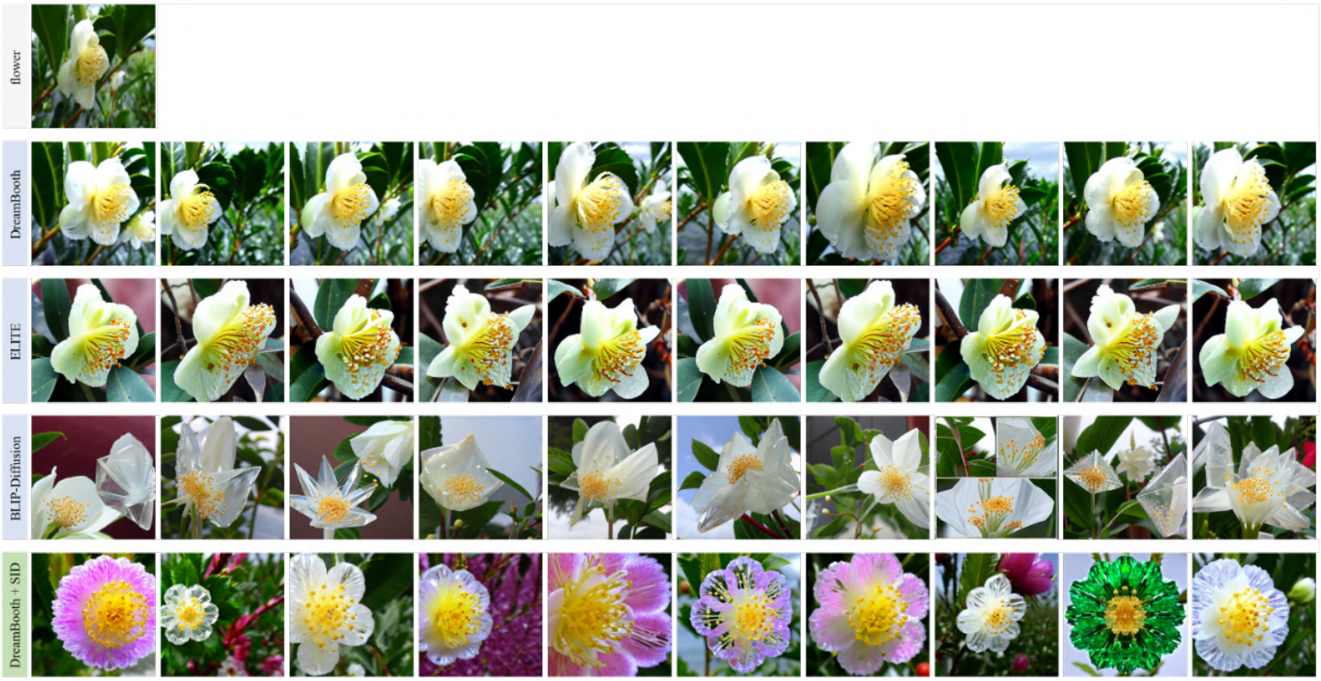


Figure E.5. (f) **Enhancement by SID for a single reference image – additional example #6.** When trained with a single reference image, DreamBooth sometimes produces results that are completely overfit to the reference image, particularly when faced with challenging generation prompts.



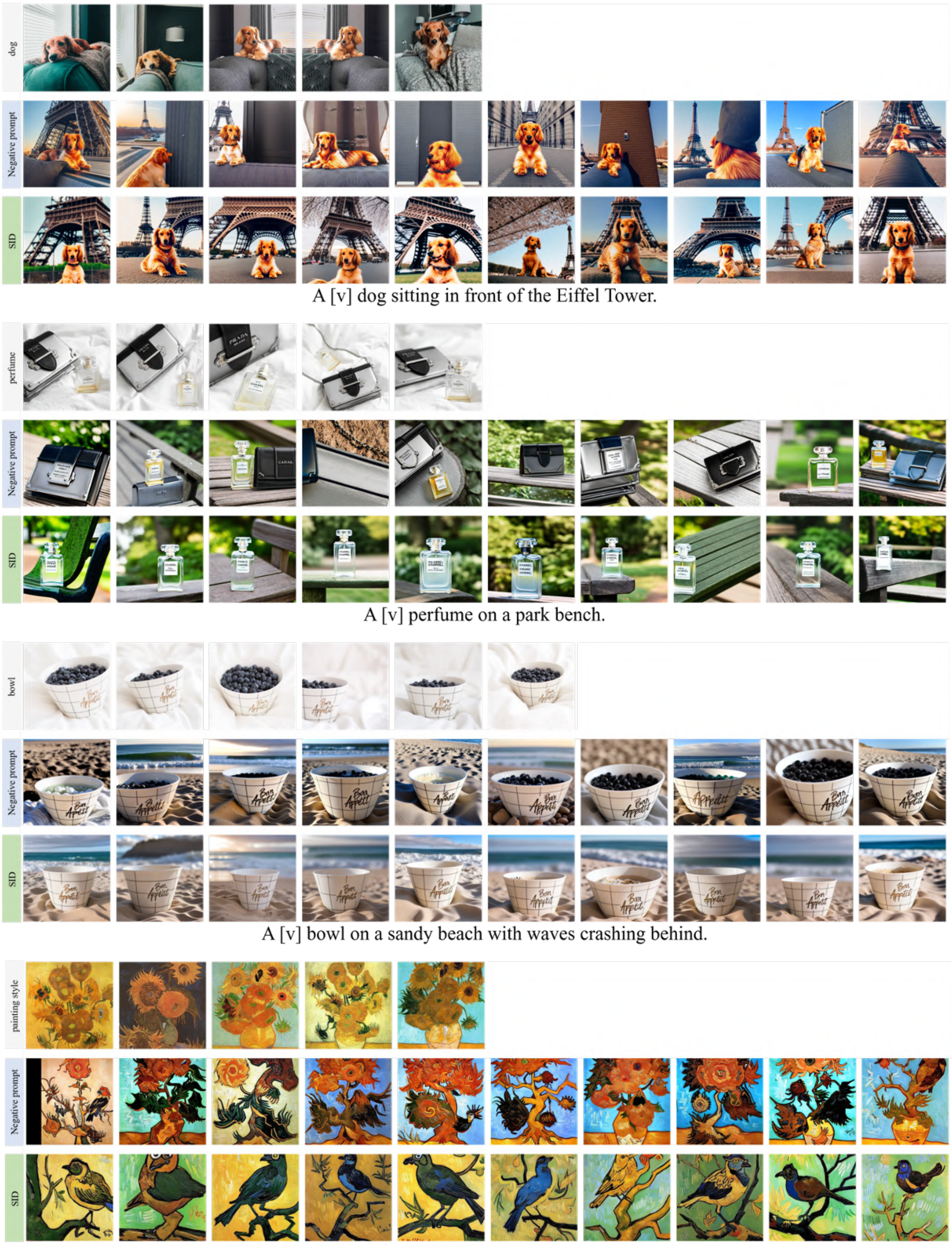
A backpack in the style of [v] toy car.

Figure E.5. (g) **Enhancement by SID for a single reference image – additional example #7.** While other models failed to generate a backpack, the SID-integrated DreamBooth successfully produced a backpack in the style of the main subject.



A [v] flower made of crystal.

Figure E.5. (h) **Enhancement by SID for a single reference image – additional example #8.** SID-integrated DreamBooth successfully altered the material of the main subject while preserving its identity.



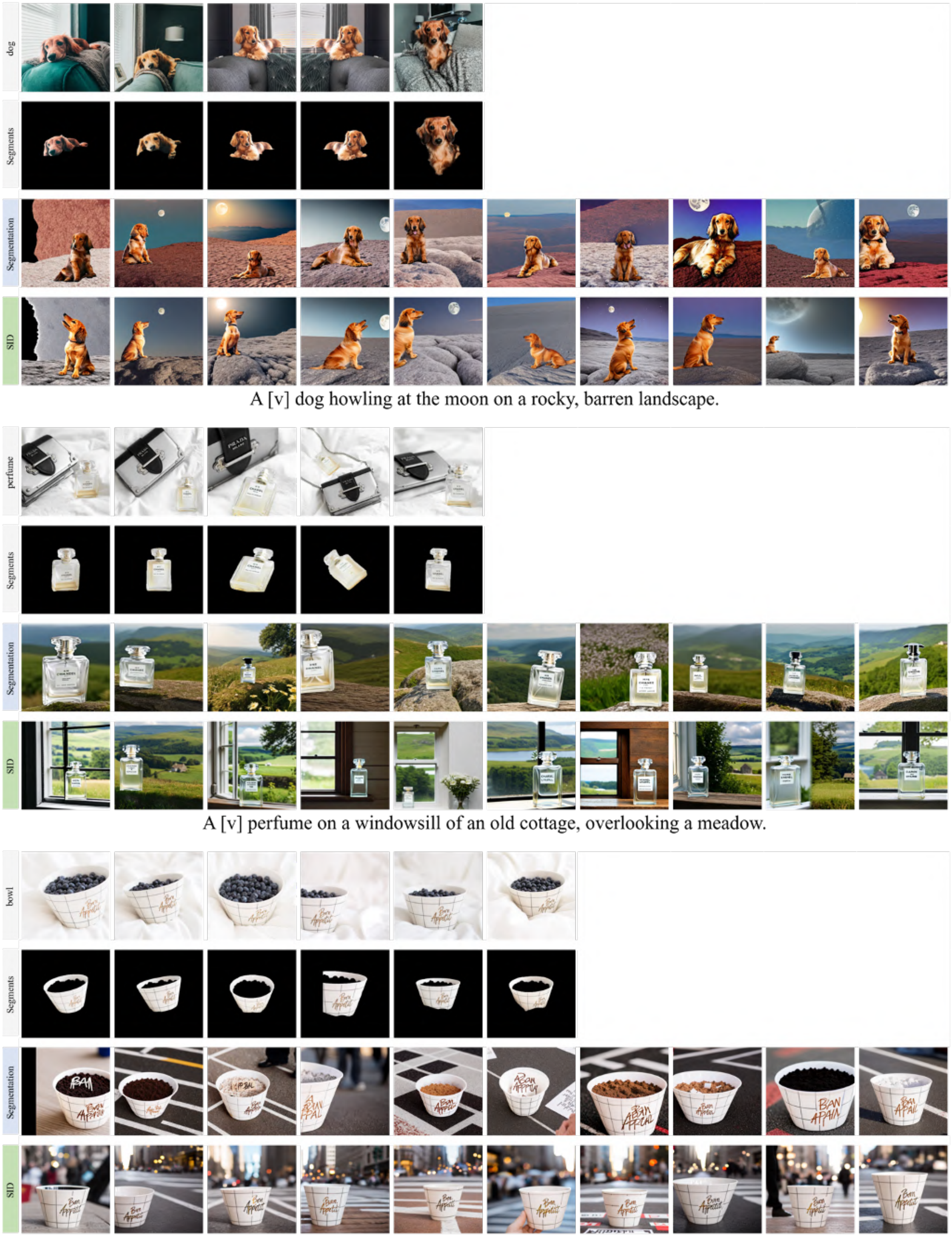
A [v] dog sitting in front of the Eiffel Tower.

A [v] perfume on a park bench.

A [v] bowl on a sandy beach with waves crashing behind.

A painting of a bird on the tree in the style of [v] art.

Figure E.6. (a) **Comparison with negative prompt – additional examples.** Even when a negative prompt is employed, it appears challenging to counter the effects of undesired embedding entanglements. Negative prompts used: "sitting on the fluffy blanket and green couch" (1st row), "next to a black Prada purse with silver hardware on a bed of white sheets" (2nd row), "full of blueberries is on the bed of white sheet" (3rd row), "a bouquet of sunflowers in the round vase" (last row). DreamBooth is used as the base model.

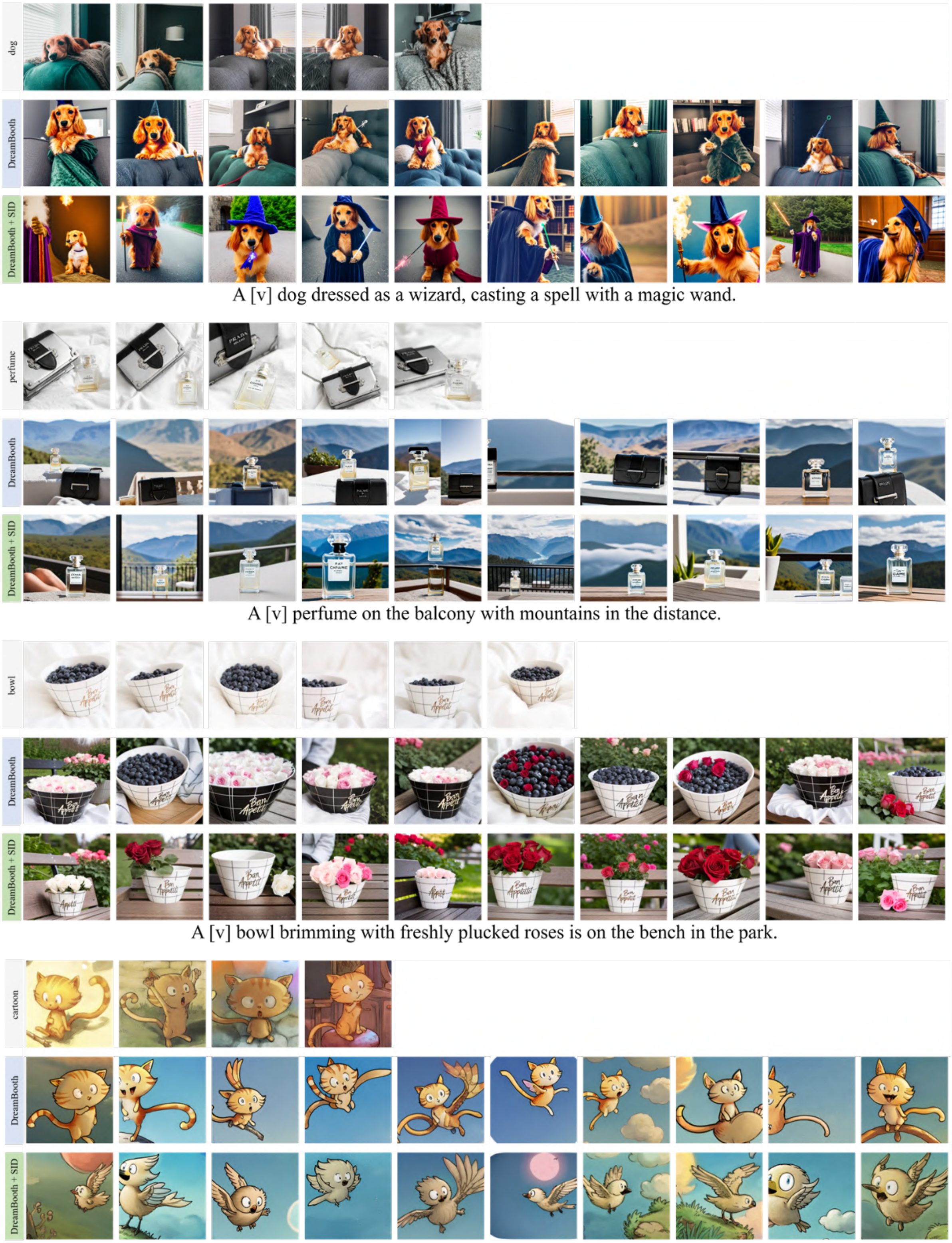


A [v] dog howling at the moon on a rocky, barren landscape.

A [v] perfume on a windowsill of an old cottage, overlooking a meadow.

A [v] bowl in the middle of a bustling city intersection.

Figure E.6. (b) **Comparison with segmentation – additional examples.** Employing segmentation to mitigate undesired embedding entanglement still presents certain limitations. The first row highlights the constraint of dynamically changing poses. The second row underscores the incapacity to generate intricate backgrounds, possibly due to the common presence of a black background in segmentation. The last row illustrates that removal of tied objects (blueberries in this case) may lead to generated images with peculiar artifacts. DreamBooth is used as the base model.



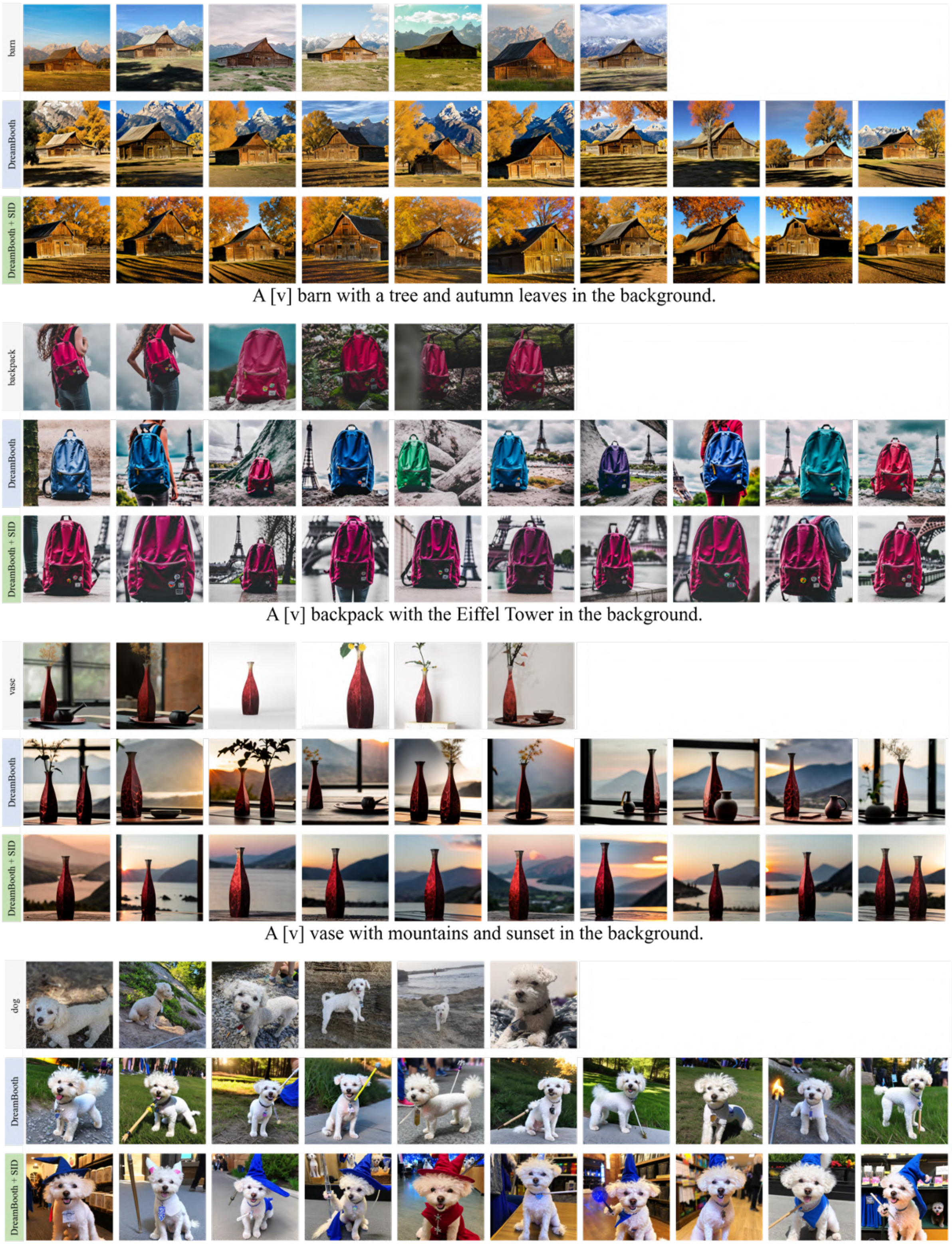
A [v] dog dressed as a wizard, casting a spell with a magic wand.

A [v] perfume on the balcony with mountains in the distance.

A [v] bowl brimming with freshly plucked roses is on the bench in the park.

A cartoon of a bird flying the sky in the style of [v] art.

Figure E.7. (a) **Comparing DreamBooth with its SID-integration in highly-biased scenarios.** DreamBooth suffers from undesired embedding entanglement represented by indoor background (1st row), nearby purse (2nd row), filled-in blueberries (3rd row), and cat substance (last row). SID-integration is definitely required for high-quality text-to-image personalization. Image credit: David Revoy (last row).



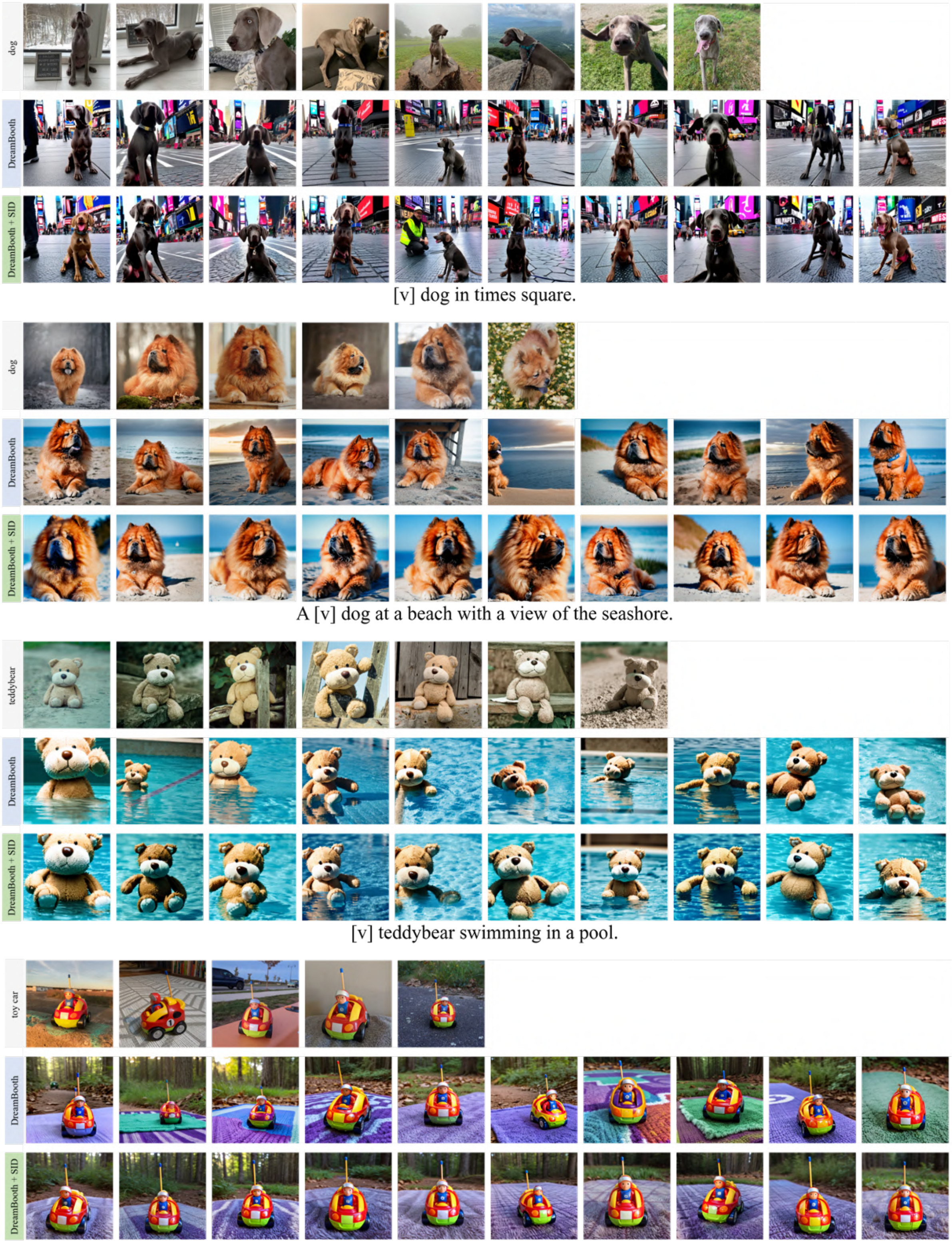
A [v] barn with a tree and autumn leaves in the background.

A [v] backpack with the Eiffel Tower in the background.

A [v] vase with mountains and sunset in the background.

A [v] dog dressed as a wizard, casting a spell with a magic wand.

Figure E.7. (b) **Comparing DreamBooth with its SID-integration in moderately-biased scenarios.** DreamBooth still suffers from undesired embedding entanglement even in moderately-biased scenarios. This is evident in the background mountain (1st row), rocky surface (2nd row), nearby teapot, saucer, and filled-in plants (3rd row), and grassy field (last row). In particular, in the second row, DreamBooth encounters difficulty preserving the identity of the subject, as it tends to change its color. SID-integration is definitely required for high-quality text-to-image personalization.



[v] dog in times square.

A [v] dog at a beach with a view of the seashore.

[v] teddybear swimming in a pool.

A [v] toy car on top of a purple rug in a forest.

Figure E.7. (c) **Comparing DreamBooth with its SID-integration in low-biased scenarios.** In scenarios with much less or almost no bias in the reference images, both DreamBooth and SID-integrated DreamBooth demonstrate remarkable performance.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[3] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023.

[4] Haotian Liu. liuhaotian/llava-v1-0719-336px-lora-merge-vicuna-13b-v1.3. https://huggingface.co/liuhaotian/llava-v1-0719-336px-lora-merge-vicuna-13b-v1.3, 2023.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[7] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv preprint arXiv:2305.14720*, 2023.

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[10] Makoto Shing. svdiff-pytorch. https://github.com/mkshing/svdiff-pytorch, 2023.

[11] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2023.

[12] OpenAI. Gpt-4 technical report, 2023.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[15] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 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*, pages 22500–22510, 2023.

[16] Salesforce. LAVIS. https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion, 2023.

[17] Salesforce. Salesforce/blip2-opt-2.7b. https://huggingface.co/Salesforce/blip2-opt-2.7b, 2023.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[19] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models.

[20] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023.

[21] Yuxiang Wei. ELITE. https://github.com/csyxwei/ELITE, 2023.