LoRA.rar: Learning to Merge LoRAs via Hypernetworks for Subject-Style Conditioned Image Generation

Supplementary Material

This document includes additional material that was not possible to include in the main paper. Sec. A1 presents additional details regarding both MLLM-based and human evaluation, further information on image generation prompts, and it also includes dataset attribution and partitioning details. Sec. A2 shows additional results: performance via standard metrics, a thorough ablation study on the hypernetwork design, results on a lightweight diffusion model, generalization to new concepts, new splits, and recontextualization output generations. Sec. A3 outlines limitations of our approach and discusses its societal impact.

A1. Additional Details

A1.1. MLLM-based Evaluation

Evaluation Prompts. We show the prompts we have used with our MLLM-based MARS^2 metric using the LLaVA-Critic-7b model. The subject assessment prompt is shown in Fig. A1, while the style assessment prompt is in Fig. A2.

Subject Assessment Prompt

System Prompt
You are a helpful assistant.

User Prompt
Your task is to identify if the test image shows the same subject as the support image.

Support image:

`{Image}`

Test image:

`{Image}`

Pay attention to the details of the subject, it should for example have the same color. However, the general style of the image may be different.
Does the test image show the same subject as the support image?
Answer with **Yes** or **No** only.

Figure A1. **Subject Assessment Prompt.** Prompt used to evaluate the subject fidelity on generated images via our MLLM-based metric MARS^2.

Style Assessment Prompt

System Prompt
You are a helpful assistant.

User Prompt
Your task is to identify if the test image shows the subject in {style} style. An example image in the {style} style is provided.

Example image in the {style} style:

`{Image}`

Test image:

`{Image}`

The example image shows an illustration of the {style} style and the details of the subject are expected to be different.
Do not check similarity with the subject.
Is the test image in the {style} style?
Answer with **Yes** or **No** only.

Figure A2. **Style Assessment Prompt.** Prompt used to evaluate the style on generated images via our MLLM-based metric MARS^2.

We test separately for correctness of the generated subject and style as we have found such approach to be more robust. We have also manually checked how accurate the MLLM model is in assessing the correctness of the subject and style, taken singularly, and found the quality to be suitable for the task. We show examples of how the MLLM judge assesses various generated images in terms of the subject or style in Fig. A3. In the first and second row, the generated images reproduce the reference subject in the reference style and, therefore, are correctly accepted by the MLLM judge. Images in third and fifth rows reproduce a generic cat (*e.g.*, white rather than gray) in the correct style, hence the MLLM judge accepts the style but not the subject preservation. The teapot in the fourth row is preserved in the generated image, but the style is incorrect (*e.g.*, more similar to an oil painting rather than watercolor painting).

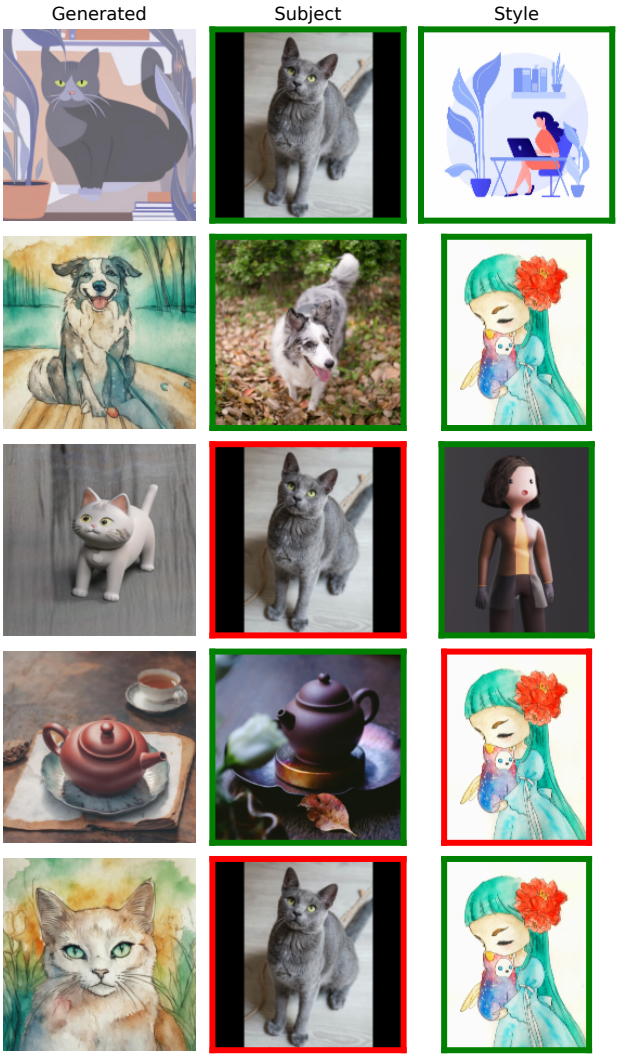

Figure A3. **MLLM Judge Assessment Samples.** This figure illustrates how the MLLM judge evaluates generated images for subject and style alignment. First column: examples of generated images. Second and third columns: reference subject and style, respectively. Green boxes indicate that the MLLM judge confirms the generated image aligns with the reference subject or style, whereas red boxes denote a mismatch.

A1.2. Human Evaluation Study

As part of the human evaluation study, we asked 25 participants to compare two generated images at a time, given reference subject and style images. The images are generated by either our approach or ZipLoRA, and they are randomly ordered in each pair. We test 25 subject-style combinations with one pair of generated images for each. The combinations are also randomly ordered. We consider two scenarios, one where we use randomly generated images and one where we take the "best" images as judged by the MLLM judge. In the "best" scenario, we gathered all the images that satisfied both subject and style according to the MLLM judge and then selected one randomly among those–there was always at least one such example for each approach.

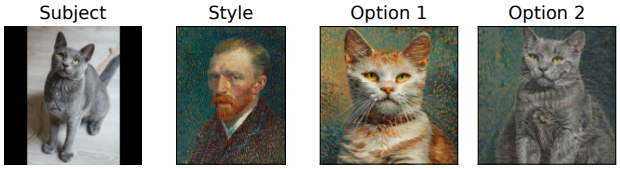

Figure A4. **Example Case for Evaluators.** Example used to teach human evaluators how to evaluate the generated images. In this example, the participant should select *Option 2* as better, because the generated image in Option 2 represents the target subject in the target style. Option 1 follows the style, but generates a random cat instead.

We introduced and explained the task to the evaluators via the example shown in Fig. A4 and the following textual instruction: *"Your task is to evaluate which of two generated images better represents the given subject and style – or if they are similarly good. You are provided with an image showing the subject (e.g. black cat) and an image showing the image style (e.g. van Gogh style painting), and two generated images such as in the example below. In this example you would select option 2 as better because it shows a cat that looks like the one in the subject image, and both images follow the style."*.

The evaluation was done via a web app that shows the images and lets the participant click on a button saying which option is better among: *"Option 1", "Similar", "Option 2"*.

A1.3. Additional Experimental Details

Prompts Used for Image Generation. The prompts used to generate the images for the main paper qualitative and quantitative results are of the form: "`A [c] <class name> in [s] style`". For "`[c]`" we used the rare token used to train the content LoRAs and for "`<class name>`" we used the same name as DreamBooth [3]. Finally, for "`[s]`" we used the short text description as in StyleDrop, in particular it corresponds to the style name that we assigned (after removing the number, if present). The full list of names is detailed in Sec. A1.4.

Additional Implementation Details. Base LoRAs are trained as in [4], for 1000 fine-tuning steps, with batch size 1, a learning rate of 5×10^{-5} and a rank of 64. The text encoder remains frozen during training. The hypernetwork used is a two-layer MLP with two separate input layers of size 1280 and 2560, followed by a ReLU activation function, a shared hidden layer of size 128, and two outputs. We train our hypernetwork for 100 different $\{L_c, L_s\}$ combinations (totalling 5000 steps), with $\lambda = 0.01$, learning rate 0.01 and the AdamW optimizer. For ZipLoRA, we use a training

	Contents	Styles
Train	backpack, backpack dog, berry bowl, candle, cat #1, colorful sneaker, dog #1, dog #5, dog #6, dog #7, duck toy, fancy boot, grey sloth plushie, monster toy, pink sunglasses, poop emoji, rc car, red cartoon, robot toy, shiny sneaker, vase	3D rendering #1, 3D rendering #3, abstract rainbow, black statue, cartoon line drawing, flat cartoon illustration #1, glowing 3D rendering, kid crayon drawing, line drawing, melting golden rendering, oil painting #3, sticker, watercolor painting #2, watercolor painting #4, watercolor painting #5, watercolor painting #6, watercolor painting #7, wooden sculpture
Validation	dog #2, dog #3, clock, bear plushie	3D rendering #2, oil painting #1, watercolor painting #1
Test	dog #8, cat #2, wolf plushie, teapot, can	3D rendering #4, oil painting #2, watercolor painting #3, flat cartoon illustration #2, glowing

Table A1. **Dataset partitioning.** Contents and styles LoRAs train/validation/test splits.

setup of 100 steps with the same λ and learning rate. The DARE, TIES, and DARE-TIES baselines are evaluated with uniform weights and a density of 0.5. For joint training, we used a multi-concept variant of Dreambooth LoRA as in [4]. In all experiments, 50 diffusion inference steps are used.

A1.4. Additional Dataset Details

Figure A5. **Test Set Samples.** Subject and styles of the test set in our data partitioning.

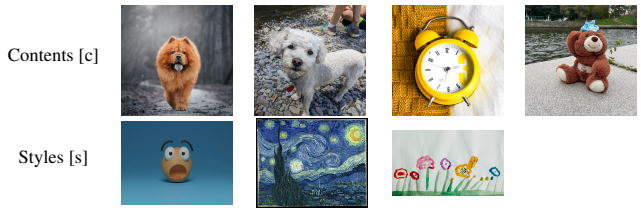

Figure A6. **Validation Set Samples.** Subject and styles of the validation set in our data partitioning.

We use the style images from the datasets collected by StyleDrop / ZipLoRA [4, 5], while the subject images are taken from the DreamBooth [3] dataset. Note that these datasets do not contain any human subjects data or personally identifiable information. We provide image attributions below for each image that we used in our experiments. We refer readers to manuscripts and project websites of StyleDrop, ZipLoRA and DreamBooth for more detailed information about the usage policy and licensing of these images.

Attribution for Style Reference Images StyleDrop project webpage provides the image attribution information here. In particular, we used the following 20 styles: S1 (3D rendering #1), S2 (watercolor painting #1), S3 (3D rendering #3), S4 (sticker), S5 (flat cartoon illustration #2), S6 (watercolor painting #5), S7 (flat cartoon illustration #1), S8 (melting golden rendering), S9 (kid crawyon drawing), S10 (wooden sculpture), S11 (oil painting #3), S12 (watercolor painting #7), S13 (watercolor painting #6), S14 (oil painting #1), S15 (line drawing), S16 (oil painting #2), S17 (abstract rainbow colored flowing smoke wave design), S18 (glowing), S19 (glowing 3D rendering), S20 (3D rendering #4). Additionally, we also used 6 styles from ZipLoRA (linked as hyperlinks): S21 (3D rendering #2), S22 (watercolor painting #2), S23 (watercolor painting #3), S24 (watercolor painting #4), S25 (cartoon line drawing), S26 (black statue).

Attribution for Subject Reference Images The DreamBooth project webpage provides the image attribution information here. Specifically, the sources of the content images that we used in our experiments are as follows (linked as hyperlinks): C1, C2, C3, C4, C5, C6, C7, C8, C9, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19, C20, C21, C22, C23, C24, C25, C26, C27, C28, C29, C30.

Dataset Partitioning There are 30 subjects and 26 styles overall. We split the subjects and styles randomly, but with the constraint that there is a good representation of different subjects and styles in each split as some subjects and styles are similar to each other. For example we aimed at avoiding only testing on different dogs or only on painting styles.

We split the subjects and styles into training, validation and test splits as shown in Tab. A1. In Fig. A5 and Fig. A6 we show images taken from the test and validation sets respectively (used to train the test and validation LoRAs).

A2. Additional Results

A2.1. Performance via Standard Metrics

Standard metrics evaluations (DINO, CLIP-I, CLIP-T) are reported in Table A2. We include this analysis for informational purposes only. As explained in Sec. 4 of the main paper, these metrics are not optimal for the joint subject-style personalization task. Specifically, DINO (CLIP-I) is maximized when the subject (style) reference images are copied without meaningful integration, so more attention should be given to MLLM and human evaluation results.

	CLIP-I	DINO	CLIP-T
Joint Training [3]	0.623	0.764	0.329
Direct Merge [6]	0.657	0.747	0.305
DARE [8]	0.630	0.576	0.360
TIES [7]	0.620	0.592	0.358
DARE-TIES [1]	0.618	0.559	0.355
ZipLoRA [4]	0.643	0.741	0.334
LoRA.rar (ours)	0.656	0.643	0.344

Table A2. **Standard Metrics.** LoRA.rar attains similar results, but these metrics are inadequate for joint subject-style changes.

A2.2. MLLM Results per Subject and Style

We provide results of MLLM evaluation for each test subject and style in Fig. A7. We report the results for both the average case as well as the best case. The results indicate that there are certain subjects and styles that are more challenging than others, for example the *can* subject or the *glowing* style. We also see that LoRA.rar and ZipLoRA are in general significantly more successful than the other approaches, and they can be successful also in cases where other approaches typically fail, for example in the case of the *wolf plushie* subject.

A2.3. Ablation Study on Hypernetwork

We conducted an ablation study on the hypernetwork design by exhaustively exploring all possible configurations to determine which components should have their merging coefficients predicted by the hypernetwork We used the validation set and MLLM judge for this investigation, and we report the results in Table A3. We observe that the best results are obtained by *Query, Output* case that we have used; however, a few other combinations also achieve good results such as *Query, Key, Output*; *Query, Value* and *Value*.

A2.4. Results on Lightweight Diffusion Model

Fig. A8 shows qualitative results produced with KOALA 700m [2], a lightweight diffusion model, further showing that LoRA.rar could be applied to other diffusion model backbones and still outperform ZipLoRA.

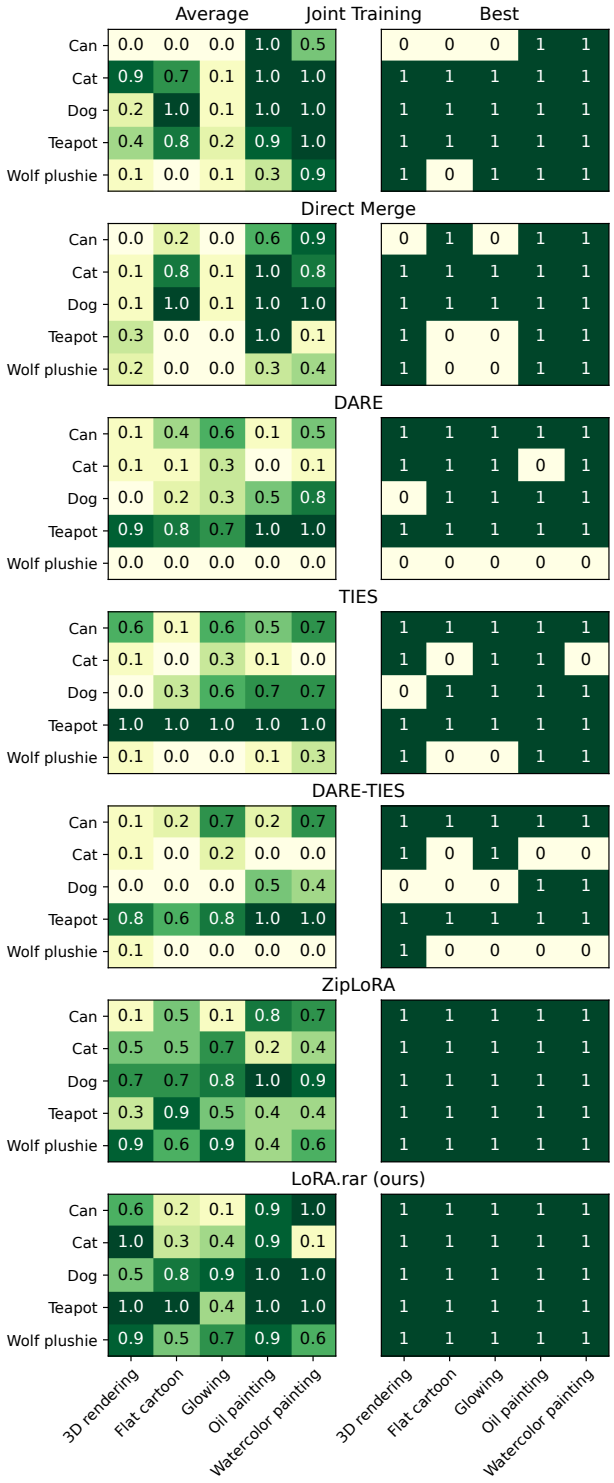

Figure A7. **MLLM Evaluation per Test Subject and Style.** Ratio of generated images with the correct content and style according to our metric MARS^2. Our solution leads to better images compared to existing approaches.

Figure A8. **Qualitative Comparison on Koala-700m.** LoRA.rar generates better images than ZipLoRA.

	Average case	Best case
Key	0.28	0.75
Value	0.43	0.83
Query	0.28	0.75
Output	0.39	0.83
Key, Value	0.40	0.92
Key, Query	0.31	0.75
Key, Output	0.44	0.75
Query, Value	0.42	0.83
Query, Output	0.48	0.92
Value, Output	0.29	0.58
Query, Key, Value	0.41	0.83
Query, Value, Output	0.23	0.33
Query, Key, Output	0.49	0.83
Key, Value, Output	0.29	0.50
Query, Key, Value, Output	0.23	0.50

Table A3. **Ablation Study via MLLM Evaluation.** Ratio of generated images with the correct content and style on the combinations of validation subjects and styles according to our MARS^{2}metric.

A2.5. Generalization to New Concepts

As common in the personalized image generation literature, we employed the DreamBooth dataset, which includes a diverse set of objects. In the main paper, we already tested generalization to new subjects (clock, teapot, and can), different from pre-training categories (see Tab. A1 for details). We consider three new furniture subjects (*toaster* collected by us; *tv*, *sofa* from the web), and a new substantially different style (*cyberpunk*). The aim of this experiment is twofold: (1) we further prove the generalization of our approach to unseen subjects-style, (2) we demonstrate the simplicity of collecting new LoRAs from single images and merge them for joint subject-style personalized image generation. Fig. A9 shows that our hypernetwork generalizes well and does not need to be trained every time a new object appears. Also in this case, we outperform ZipLoRA in MARS^2 (0.8 vs. 0.6).

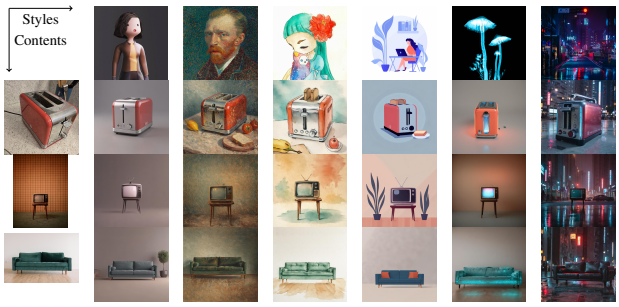

Figure A9. **Generalization to New Subjects and Styles.** LoRA.rar performs well also on new objects and styles.

A2.6. Generalization to New Splits

We re-trained the hypernetwork using 2 new splits, with the same hyperparameters as in the other experiments in the paper. The two splits that we consider are:

1. Training subjects: objects (no animals included);
   Test subjects: stuffed animals;
   Training styles: 3D renderings;
   Test styles: cartoon.
2. Training subjects: animals and stuffed animals;
   Test subjects: objects;
   Training styles: watercolor paintings;
   Test styles: abstract rainbow, wooden sculpture, melting golden rendering.

The results are shown in Fig. A10. Despite the challenging setups with no overlap between training and test set macro-categories, our method still performs well and outperforms ZipLoRA, even if the results are slightly worse than in setups with more diverse training data, as expected.

In both (1) and (2), we observe that LoRA.rar better preserves the style (*e.g.*, red-bordered images) and the subject identity (*e.g.*, blue images). At the same time, it reduces hallucinations (*e.g.*, green image, where ZipLoRA unnecessarily repeats the subject), degenerate outputs (*e.g.*, yellow image, where the subject is missing), or unrealistic samples (*e.g.*, in wood style, our samples exhibit a more wooden look and do not float in the air, unlike the first and third outputs

Figure A10. **Generalization to New Splits.** LoRA.rar performs well also when trained and tested on more challenging splits.

of ZipLoRA).

A2.7. Additional Qualitative Results

In Fig. A11 and Fig. A13 we report a recontextualization analysis for different subjects and styles, demonstrating the effectiveness of our approach.

A3. Discussion

A3.1. Limitations

Our approach exhibits certain limitations with specific subjects, particularly the *can*. This limitation is shared by the other tested model merging methods as well. The *can* subject is especially challenging because generative models struggle to accurately render text on objects (as we can see in Fig. A12).

Furthermore, we note that while the MLLM judge is useful for the task of assessing generated images in terms of content and style, it is not perfect and, for example, it may overlook small details specific to the subjects.

A3.2. Societal Impact

Our work makes it possible to generate personalized images that follow a given style and show a given subject, for example one's pet in watercolor painting style. In particular we make generating personalized images significantly more accessible than before as our solution can be deployed on smartphones, enabling real-time merging of LoRA parameters needed for the personalization. However, this brings risks that are shared with image generative models and image editing methods in general. These solutions can be used for creating deceptive content, and with our method it is even easier than before. Addressing the risks of misuse is an ongoing research priority in generative AI.

References

[1] Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's mergekit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024. 4

[2] Youngwan Lee, Kwanyong Park, Yoorhim Cho, Yong-Ju Lee, and Sung Ju Hwang. Koala: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis. In *Advances in Neural Information Processing Systems*, 2024. 4

[3] 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*, 2023. 2, 3, 4

[4] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, 2024. 2, 3, 4

[5] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, Yuan Hao, Glenn Entis, Irina Blok, and Daniel Castro Chin. Styledrop: Text-to-image synthesis of any style. In *Advances in Neural Information Processing Systems*, 2023. 3

[6] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022. 4

[7] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2024. 4

[8] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024. 4

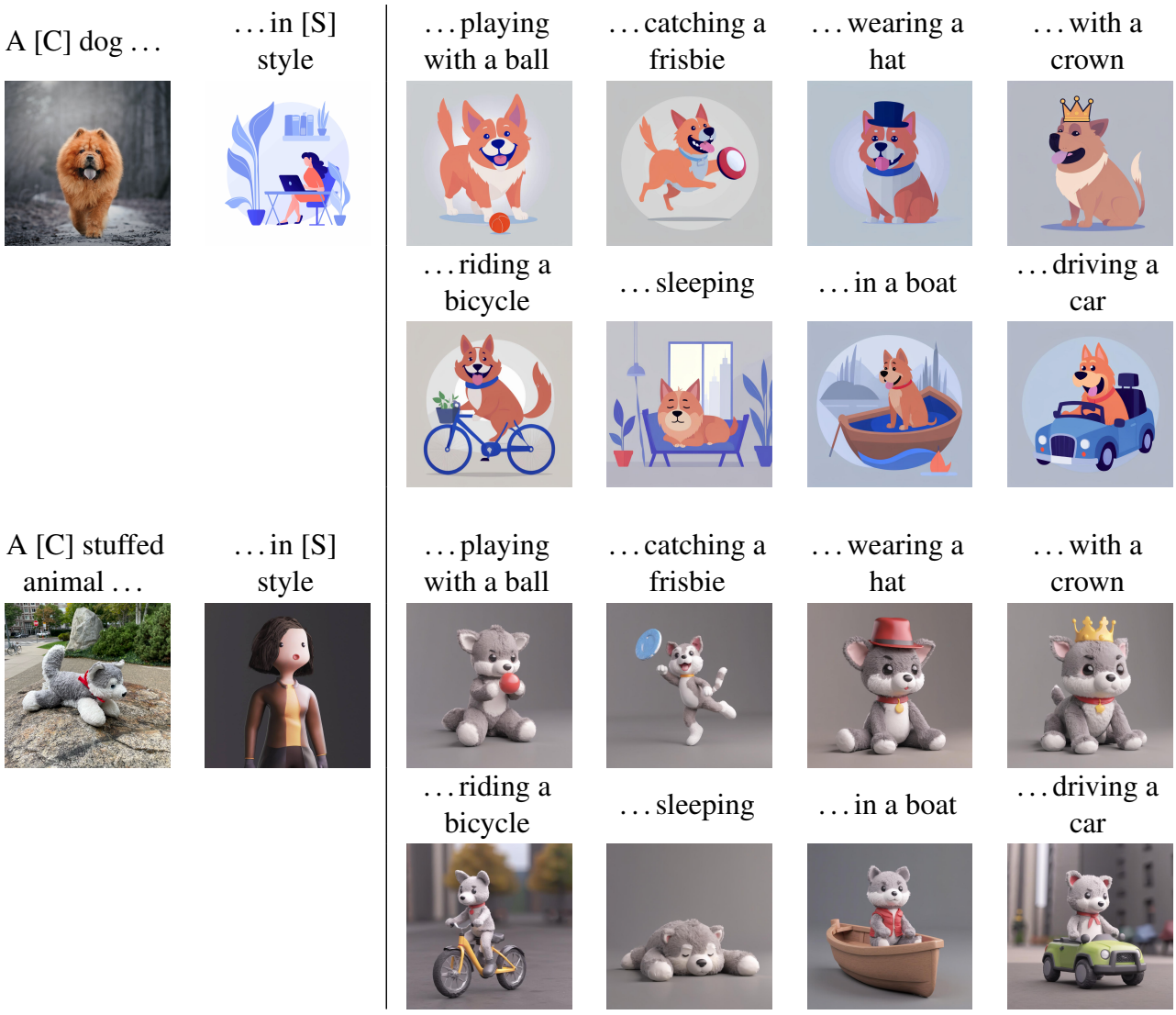

Figure A11. **Recontextualization Output Generations**. Generated outputs using various prompts for the contents "dog2" and "wolf plushie".

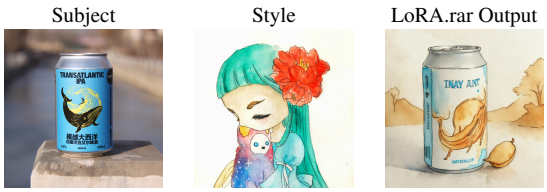

Figure A12. **Limitation Example.** Example of a challenging generation case, where the generated text and logo are not accurate.

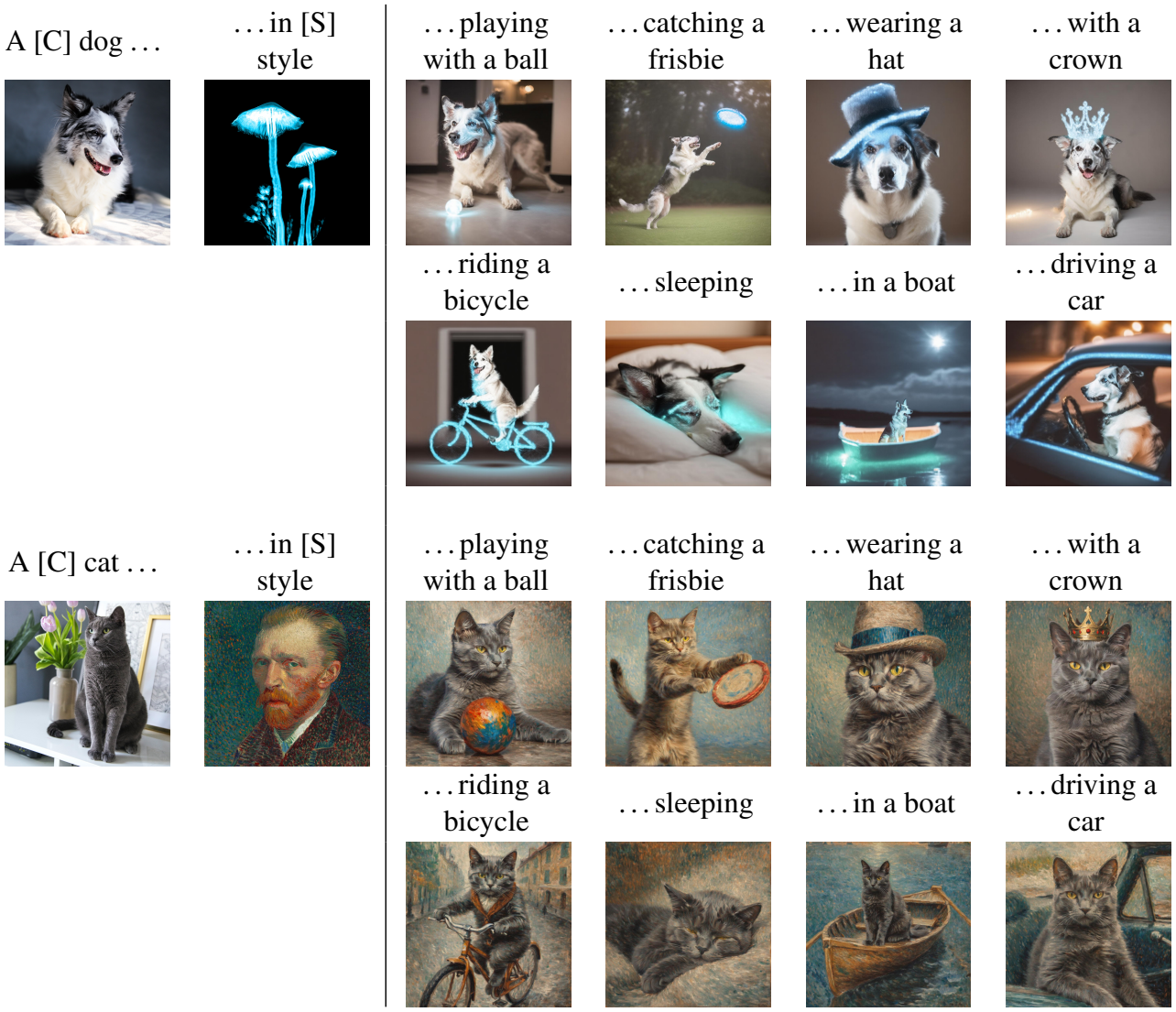

Figure A13. **Recontextualization Output Generations**. Generated outputs using various prompts for the contents "dog8" and "cat2".