

A DETAILS OF BENCHMARK DATASET

A.1 DATASETS

This study leverages two key datasets for benchmark:

- **Custom collection:** We generated custom characters such as cartoon style *cat* and *dog*, created using the *character sheet* trick ⁴ popular within the Stable Diffusion community. This set comprises 20 unique characters, where we trained a LoRA per character.
- **CustomConcept101:** We used a popular dataset Kumari et al. (2023) CustomConcept101 that includes several diverse objects such as *plushie bunny*, *flower*, and *chair*. All 101 concepts are utilized.

Leveraging the datasets above, we trained LoRAs to represent each concept, totaling to 131 LoRA models. For every competitor, the base stable diffusion model cited in the relevant paper is used. For instance, ZipLoRA Shah et al. (2023) employs SDXL, while MixOfShow Gu et al. (2023) utilizes EDLoRA alongside SDv1.5. Similarly, our method uses SDv1.5.

A.2 EXPERIMENTAL PROMPTS

To evaluate the merging capabilities of the methods, we created 200 text prompts designed to represent various scenarios such as (the corresponding LoRA models are indicated within paranthesis):

- A cat and a dog in the mountain (blackcat, browndog)
- A cat and a dog at the beach (blackcat, browndog)
- A cat and a dog in the street (blackcat, browndog)
- A cat and a dog in the forest (blackcat, browndog)
- A plushie bunny and a flower in the forest (plushie_bunny and flower_1)
- A cat and a flower on the mountain (blackcat, flower_1)
- A cat and a chair in the room (blackcat, furniture_1)
- A cat watching a garden scene intently from behind a window, eager to explore. (blackcat, scene_garden)
- A cat playfully batting at a Pikachu toy on the floor of a child’s room. (blackcat, toy_pikachu1)
- A cat cautiously approaching a plushie tortoise left on the patio. (blackcat, plushie_tortoise)
- A cat curiously inspecting a sculpture in the garden, adding to the scenery. (blackcat, scene_sculpture1)

B ADDITIONAL QUANTITATIVE ANALYSIS

In addition to the results presented in the main paper, we apply further experiments to assess the performance of our method in detail. Specifically, we apply instance segmentation methods to the composed images to identify and isolate

object instances. For this, we use SEEM (Zou et al., 2024) to segment the objects within the images. After segmentation, we calculate the similarity metrics separately for each object instance, allowing for a more granular comparison of the methods. We perform these evaluations on a set of 700 images per method, as shown in the table. The results demonstrate that our method significantly outperforms others across multiple metrics. In particular, we calculate DINO scores, which further

		Merge	Composite	ZipLoRA	Mix-of-Show	Ours
CLIP	Min.	76.0% \pm 8.7%	76.2% \pm 7.2%	73.4% \pm 8.1%	75.2% \pm 9.5%	83.3% \pm 5.5%
	Avg.	79.5% \pm 8.3%	79.7% \pm 6.8%	77.1% \pm 7.6%	78.7% \pm 9.2%	87.1% \pm 4.9%
	Max.	82.5% \pm 8.1%	82.5% \pm 6.7%	80.6% \pm 7.6%	81.7% \pm 9.2%	89.8% \pm 4.8%
DINO	Min.	37.0% \pm 15%	30.3% \pm 13%	36.9% \pm 13%	37.5% \pm 17%	47.2% \pm 14%
	Avg.	43.7% \pm 17%	38.5% \pm 13%	49.6% \pm 15%	48.0% \pm 22%	57.3% \pm 14%
	Max.	50.5% \pm 17%	49.5% \pm 14%	53.3% \pm 16%	55.6% \pm 23%	69.1% \pm 14%

⁴<https://web.archive.org/web/20231025170948/https://semicolon.dev/midjourney/how-to-make-consistent-characters>

highlight the effectiveness of our approach compared to competing methods. Moreover, we also compute CLIP scores as additional evidence of our method’s superior performance.

C ADDITIONAL QUALITATIVE RESULTS

Comparison with OMG. We perform a qualitative comparison between our method, CLoRA, and OMG (Kong et al., 2024). OMG relies on off-the-shelf segmentation methods to isolate subjects before generating images. As seen in Fig. 7, while this enables well-defined subject boundaries, the performance of OMG is heavily dependent on the accuracy of the segmentation model. Errors in segmentation can result in incomplete or incorrect generation, particularly in complex scenes involving multiple interacting subject. For instance, if the segmentation model fails to detect a flower, this may prevent the correct placement of the LoRA in the composition (see Fig. 7 bottom-left). Moreover, since OMG depends on the base image generated by the Stable Diffusion model, it also encounters the attention overlap and attribute binding issues identified by Chefer et al. (2023). For instance, if the Stable Diffusion model does not generate the required objects in the base image from the text prompt ‘A man and a bunny in the room’, then OMG cannot produce the desired composition. This issue is apparent in Fig. 7, where the rightmost image shows that the base model generated only a bunny, omitting the man. In contrast, CLoRA bypasses the need for explicit segmentation by directly updating attention maps and fusing latent representations. This ensures that each concept, represented by different LoRA models, is accurately captured and preserved during generation. The comparison in Fig. 7 demonstrates that CLoRA produces more coherent compositions, maintaining the integrity of each concept even in challenging multi-concept scenarios.

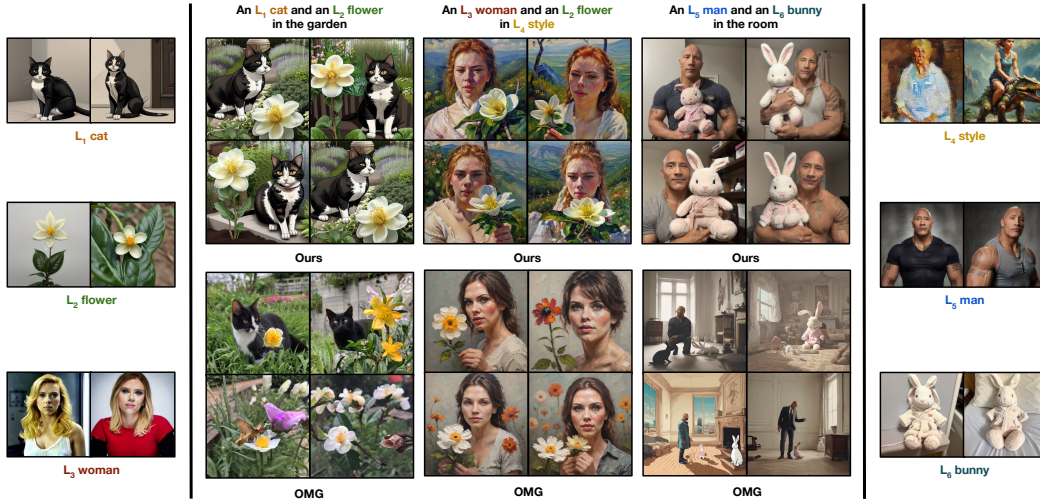


Figure 7: **Qualitative comparison with OMG.** Our method (top row) consistently produces more coherent and accurate compositions compared to OMG (bottom row). By leveraging attention map updates and latent fusion, CLoRA effectively handles multi-concept generation without relying on segmentation, leading to higher quality results, particularly in complex scenes.

Extensive Qualitative Results. The rest of the Supplementary Materials will provide additional qualitative comparisons which contain the following competitors: Mix of Show Gu et al. (2023), MultiLoRA Zhong et al. (2024), LoRA-Merge Ryu (2023), ZipLoRA Shah et al. (2023), and Custom Diffusion Kumari et al. (2023) on various LoRAs and prompts. Figure 8 compare LoRA-Merge and MultiLoRA using three combined LoRAs, while later figures expand the comparison to include all methods across two separate LoRAs.

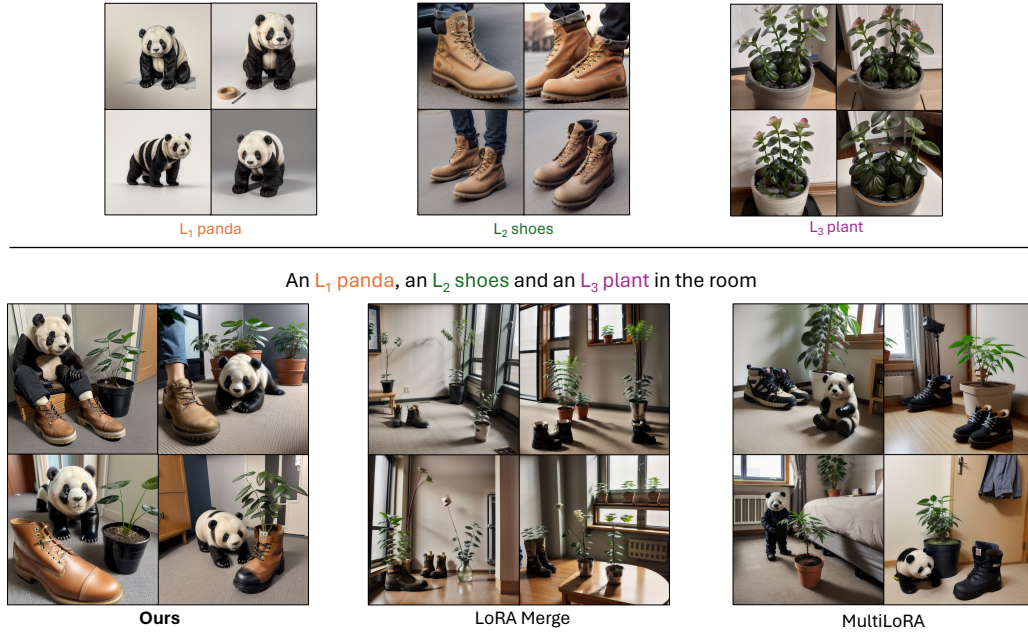


Figure 8: **Qualitative comparison of CLoRA** with other LoRA methods using 3 LoRAs to generate a single image. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

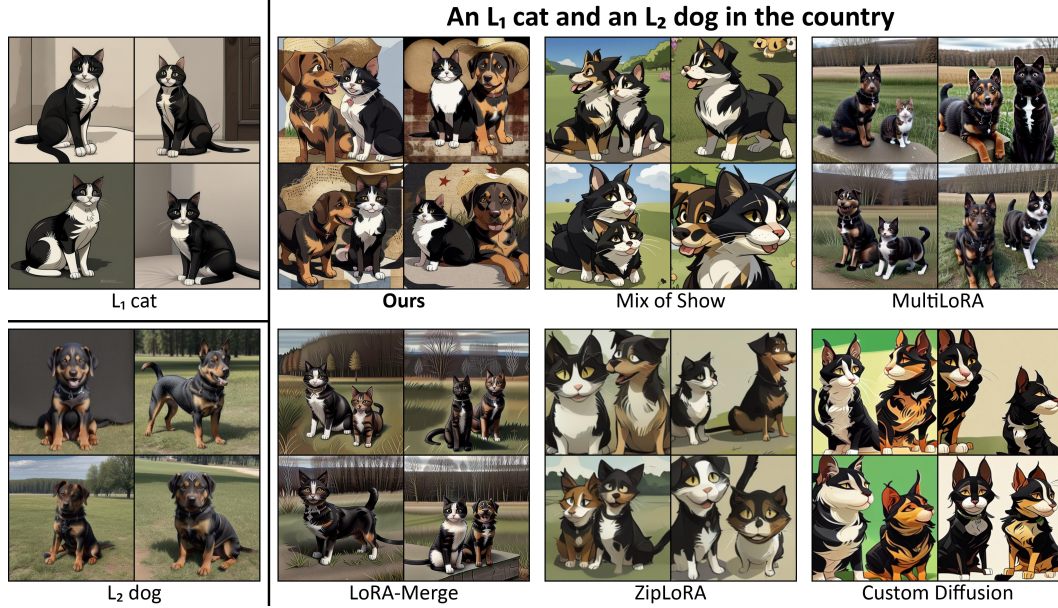


Figure 9: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.



Figure 10: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

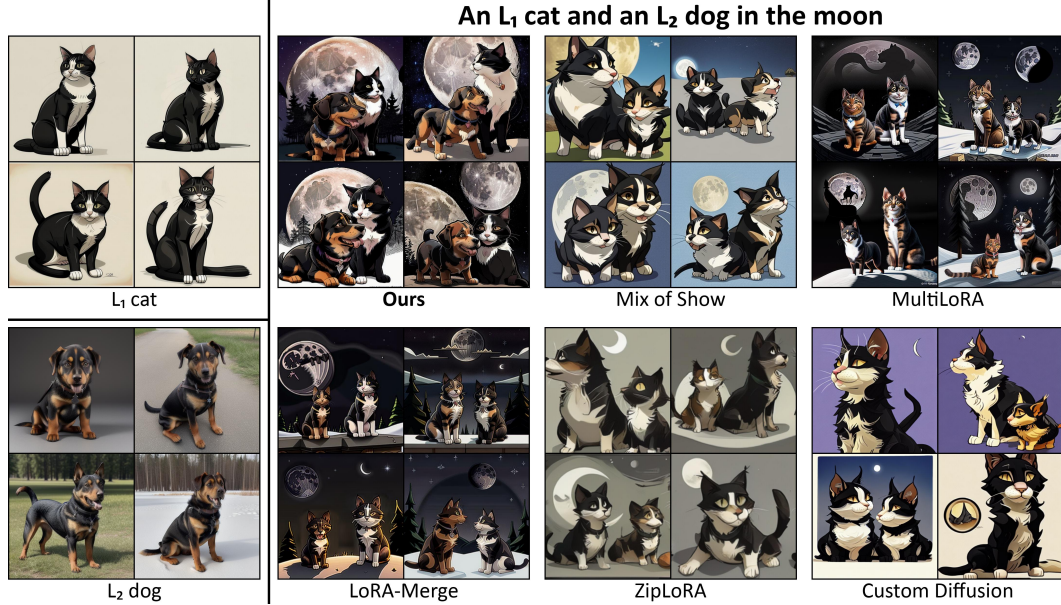


Figure 11: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

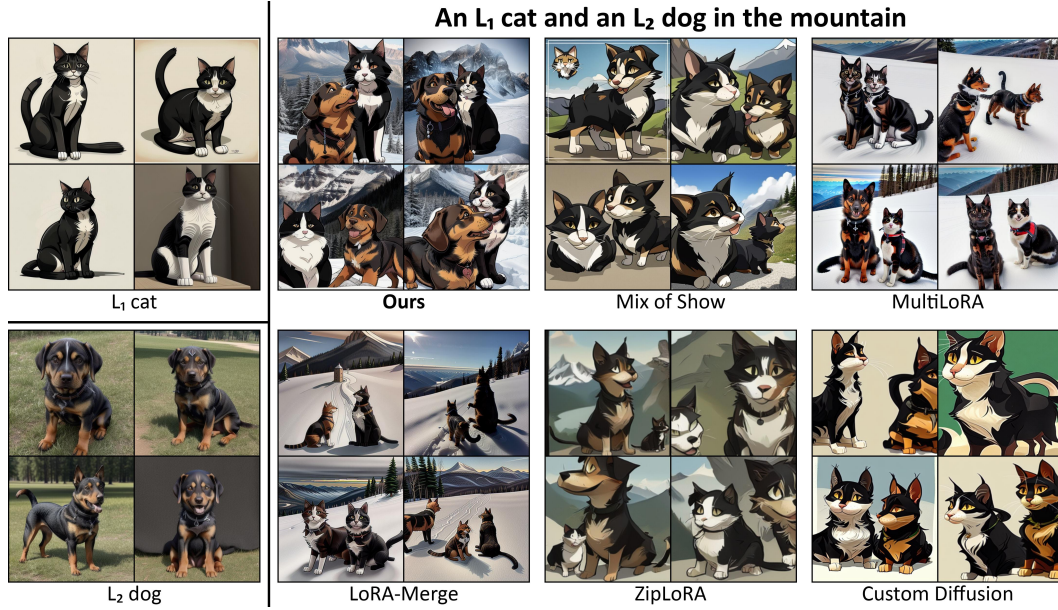


Figure 12: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.



Figure 13: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

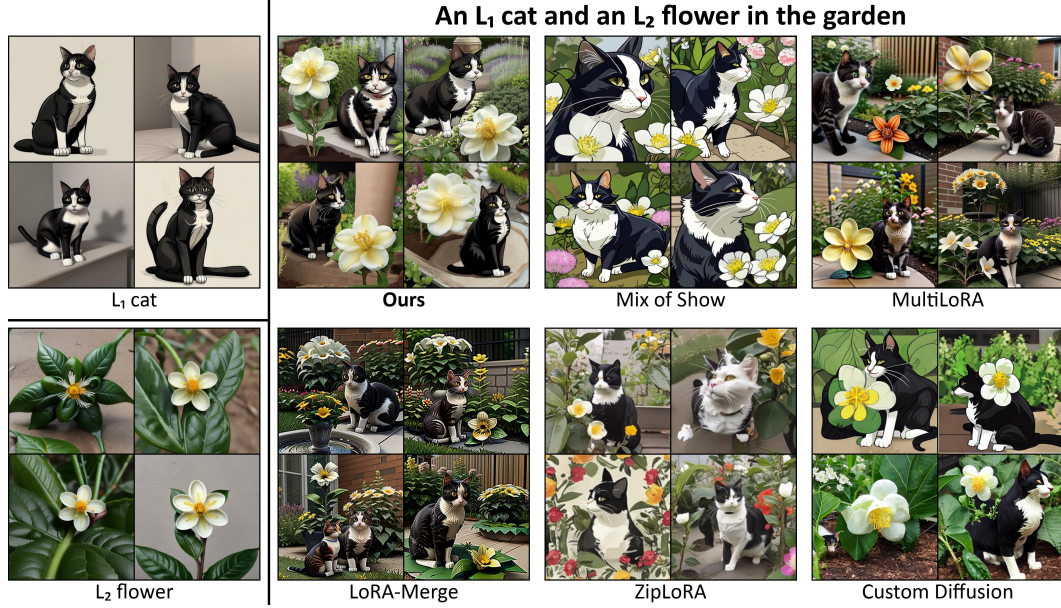


Figure 14: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

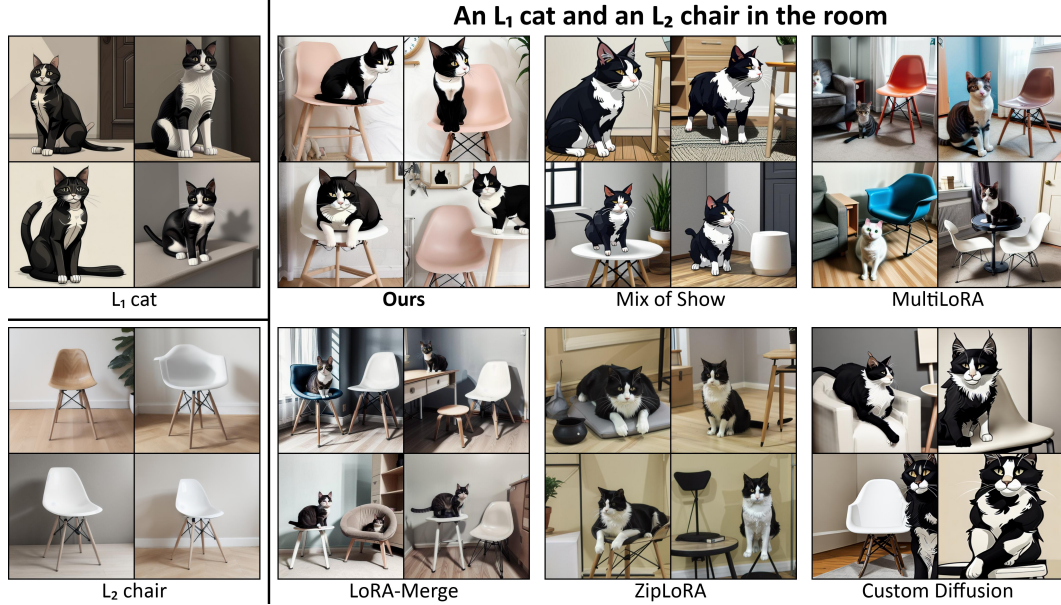


Figure 15: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

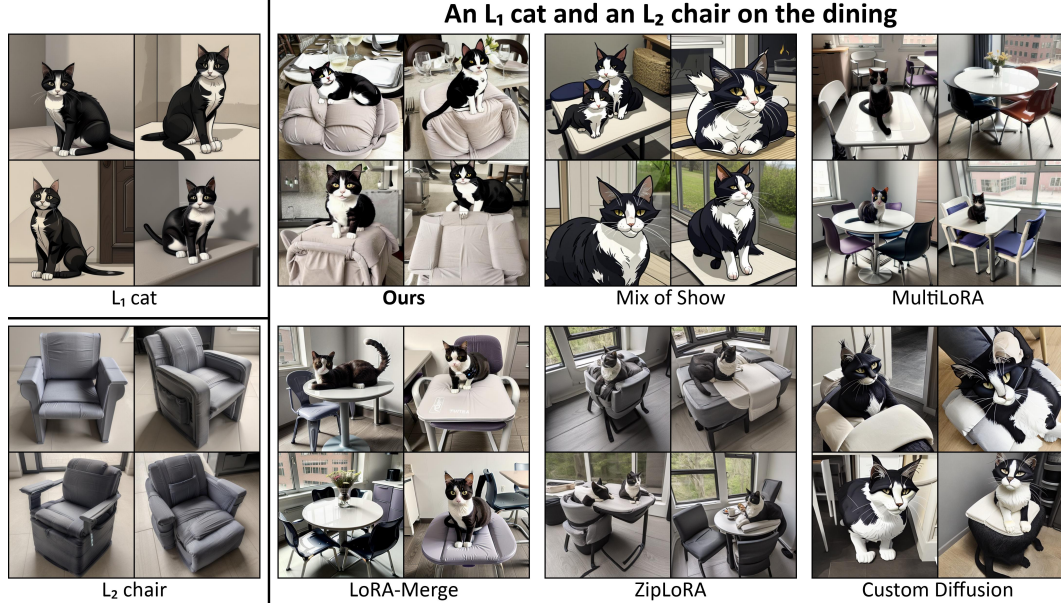


Figure 16: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.



Figure 17: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

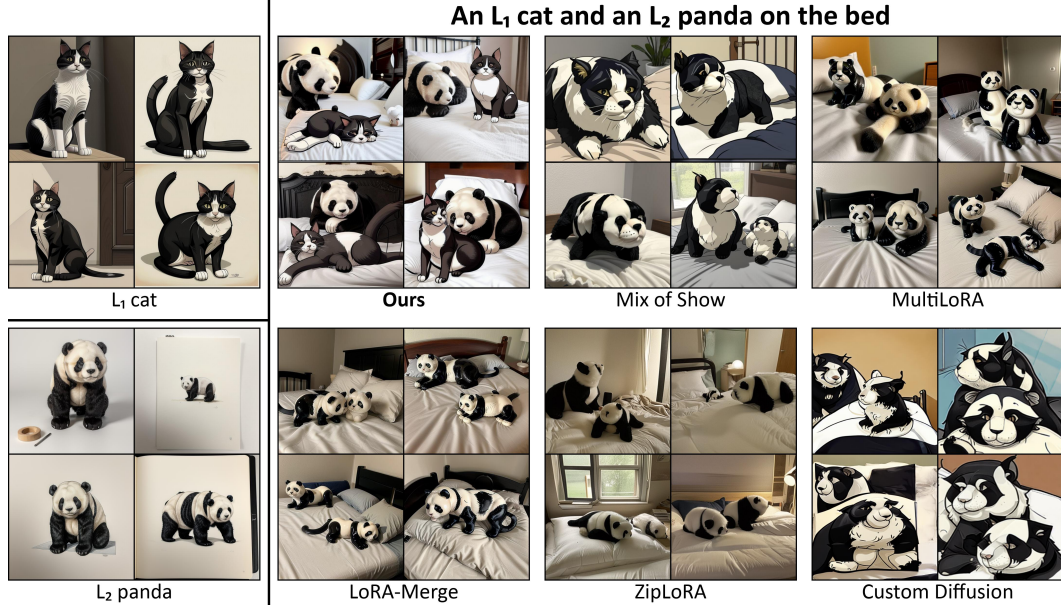


Figure 18: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

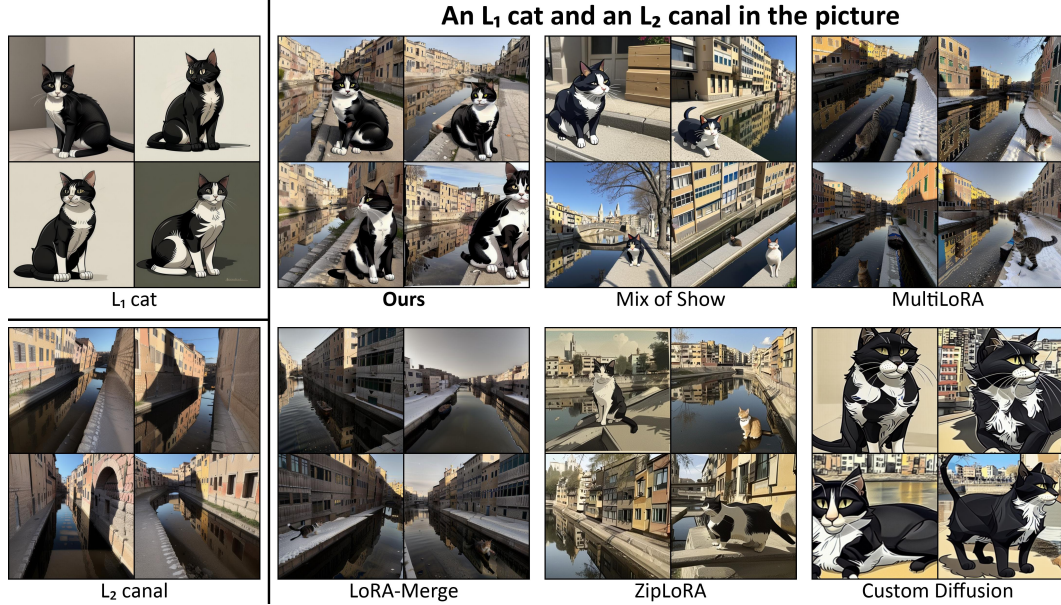


Figure 19: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

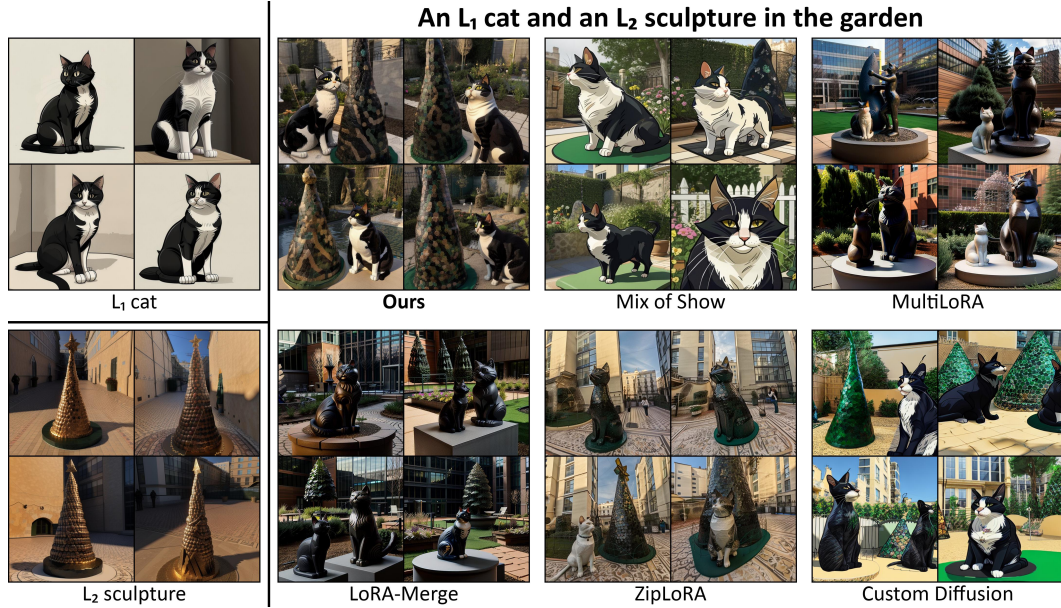


Figure 20: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

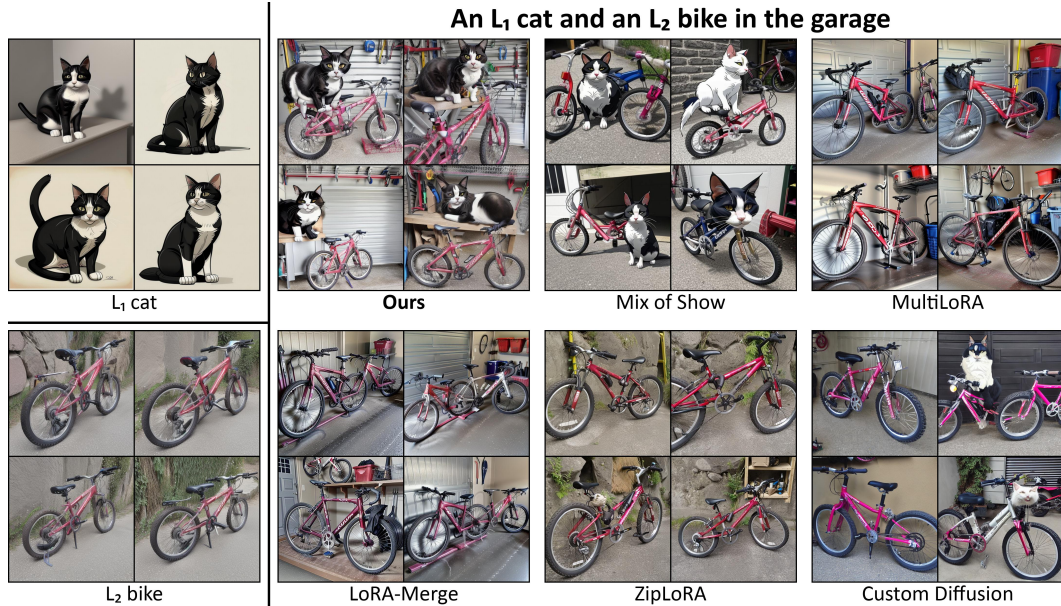


Figure 21: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.



Figure 22: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

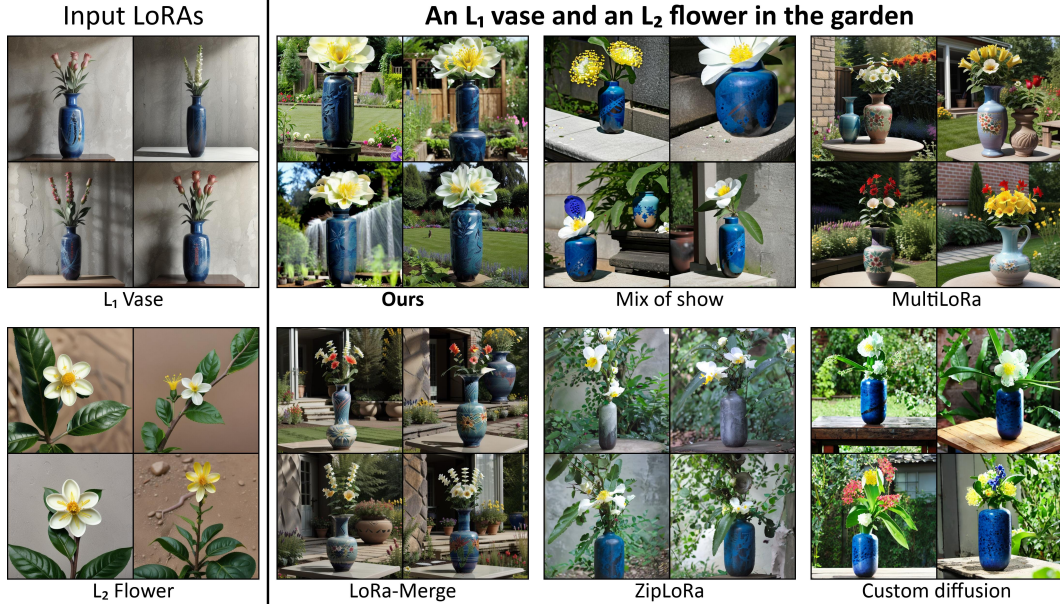


Figure 23: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.



Figure 24: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

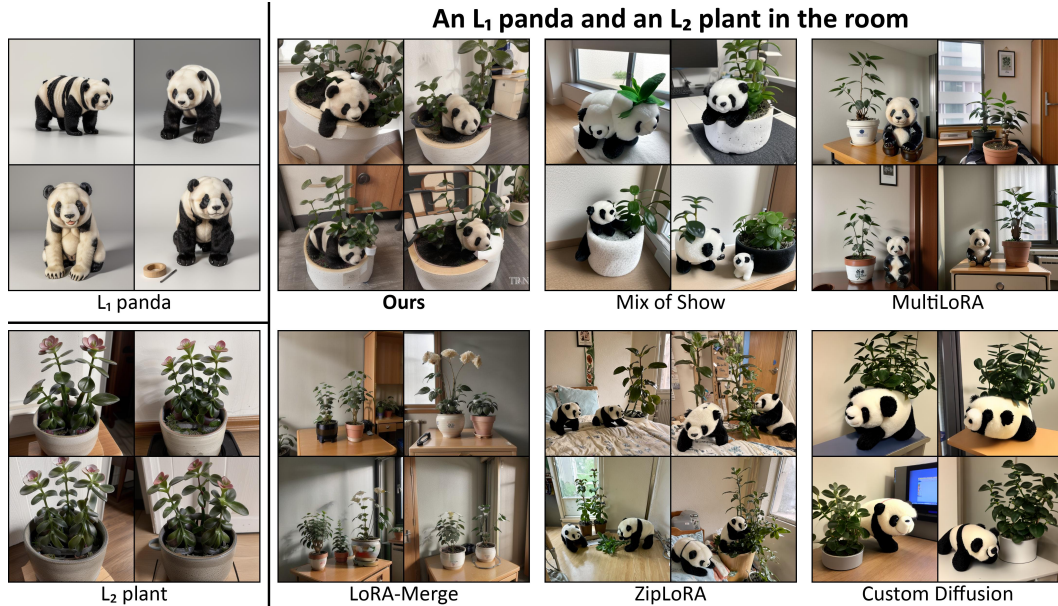


Figure 25: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

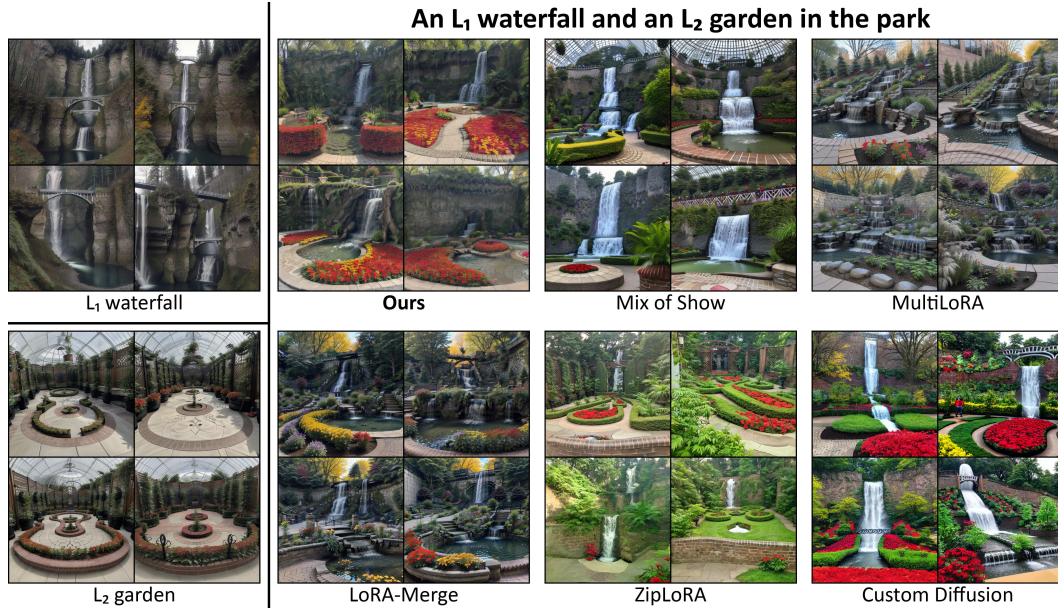


Figure 26: **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.