A. Ethics Statement & Limitations

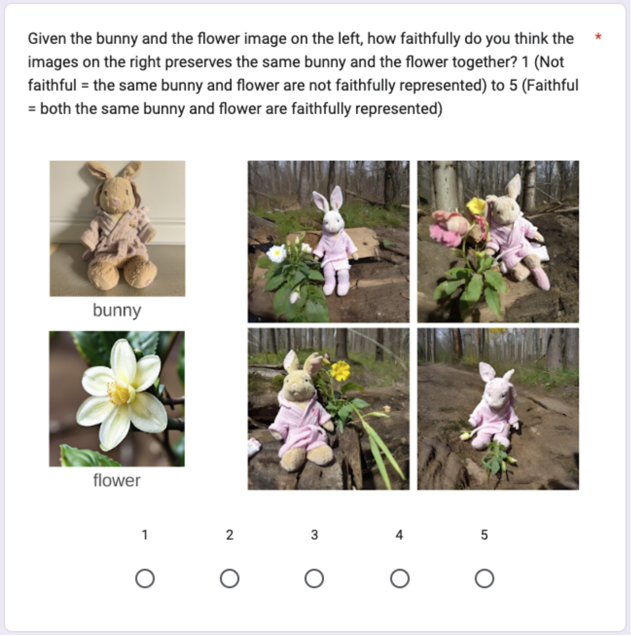
A.1. Ethics Statement

While our method democratizes creativity by simplifying the process of art creation, it also introduces ethical considerations that must be taken into account. Our method enable the generation of personalized images with minimal effort, and opens the door to transformative opportunities in art and design. However, as noted by [28], it necessitates a comprehensive and thoughtful discourse around their ethical use to prevent potential abuses. In addition to these concerns, our user study strictly adheres to anonymity protocols to safeguard participant privacy.

The capability of our method to effortlessly generate personalized images also poses risks of misuse in several harmful ways, such as the creation of deepfakes. These can be used to forge identities or manipulate public opinion, a concern underscored by Korshunov and Marcel [31].

B. User Study Details

We recruited 50 participants through the Prolific platform[3]. Each participant was shown 48 images, and asked to rate how faithfully each method preserved the concepts represented by the LoRAs (on a scale from 1 = "Not faithful" to 5 = "Very faithful"). The order of images were randomized per participant. Please see Fig. 8 to see a screenshot of our user study.



Given the bunny and the flower image on the left, how faithfully do you think the images on the right preserves the same bunny and the flower together? 1 (Not faithful = the same bunny and flower are not faithfully represented) to 5 (Faithful = both the same bunny and flower are faithfully represented) *

1	2	3	4	5
○	○	○	○	○

Figure 8. **Screenshot of our user study.** Each participant was shown images generated by LoRA models (on the left) and 4 images generated by the method (ours or competitors). Users are then requested to rate from 1-5 (Not faithful/Faithful) based on how well the generated images reflect the concepts depicted in the LoRA models.

C. Style LoRA Usage Modes

Our framework supports versatile use of style LoRAs: they can be applied globally to the complete composition or restricted to a single LoRA, according to the user's needs. This flexibility allows users to achieve different artistic effects depending on their creative intent. In Figure 9, the top row shows how model attends to each subject token.

Global Style Application: When a style LoRA is applied globally alongside all prompts, its attention mask spans the whole image and the style affects every pixel, matching standard practice. This mode is useful when users want a consistent artistic style across the entire composition. In Figure 9, middle row shows how model attends to each subject and style LoRA token when style is applied globally.

Subject-Specific Style Application: When the same style LoRA is activated together with a particular subject LoRA (e.g., pairing a sketch-style LoRA with the cat LoRA in a scene that also contains a flower), the contrastive loss confines the stylistic effect to the cat region and leaves the flower untouched. This approach enables users to apply different styles to different subjects within the same composition. In Figure 9, bottom row shows how model attends to each subject and style LoRA token when style is applied to only one subject.

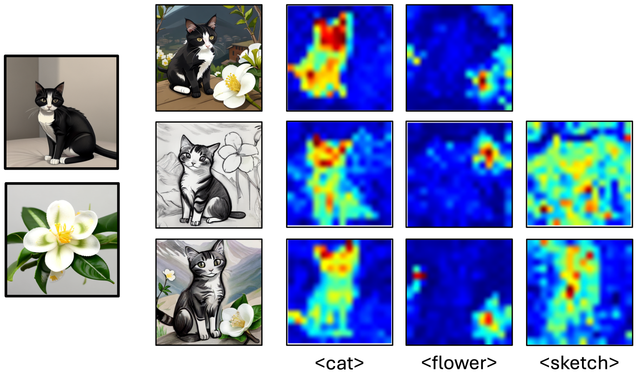


Figure 9. Attention maps corresponding to subject and style LoRAs using CLoRA

D. Additional Results

D.1. Multi-Class Object Handling

Our method is capable of handling prompts containing multiple objects from the same super-class, such as multiple people or two cats. Since we assign respective LoRAs to

[3] http://prolific.com.

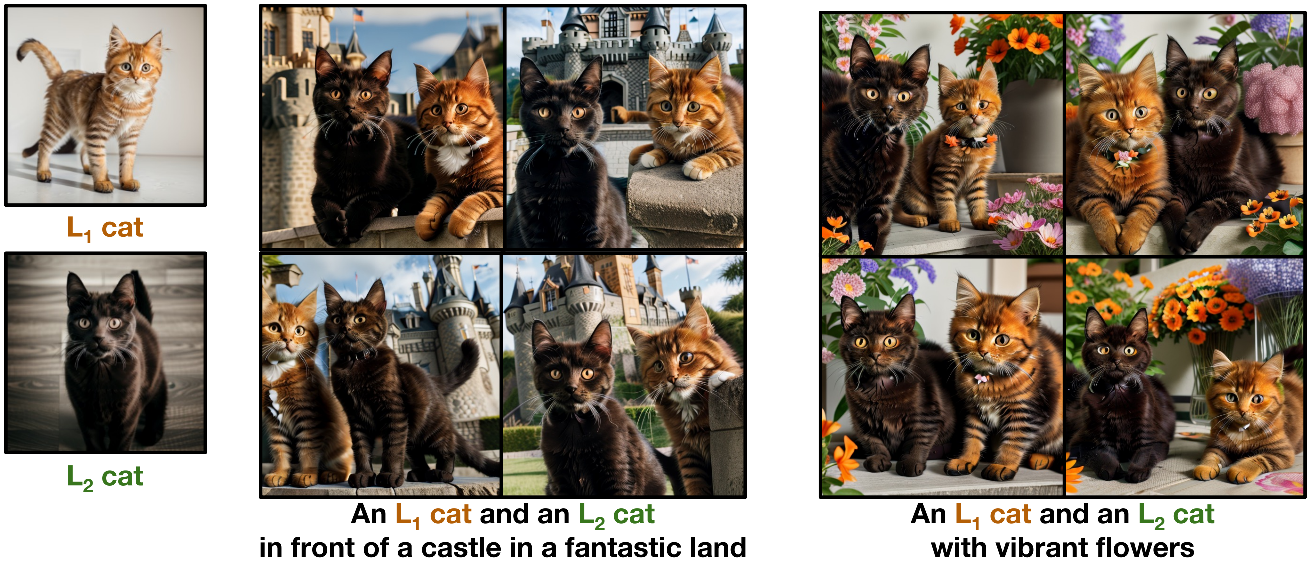

L_1 cat
L_2 cat
An L_1 cat and an L_2 cat
in front of a castle in a fantastic land
An L_1 cat and an L_2 cat
with vibrant flowers

Figure 10. Qualitative results showing that CLoRA is capable of generating images using LoRAs that has similar subjects.

individual tokens (e.g., L1 will be a positive pair with the first ”person” token, while L2 will be paired with the second ”person” token) and our separation mechanism hinges on LoRA-specific attention groups rather than coarse class labels, our method can handle such cases effectively.

This capability is demonstrated in the quantitative comparison on compositions with 3-5 LoRAs containing multiple people, where our method performs comparably to state-of-the-art methods that use special conditions to enforce separation, while our method achieves this without any auxiliary pose or conditioning inputs.

D.2. Similar Subject Compositions and Complex Scenarios

Figure 10 shows CLoRA's capabilities of generating images with similar subjects. Figure 11 showcases the CLoRA's ability to merge LoRAs in complex and interacting scenes. Our method can handle visually complex scenes that contain many objects (e.g., bottles, plates, sea on the background, or ship and ball within the same scene). These visuals show that each LoRA subject retains its unique attributes without any cross-subject leakage.

E. Details of Benchmark LoRA Collection

We propose 131 pre-trained LoRA models and 200 text-prompts for multi-LoRA composition. The details of our dataset is given below.

E.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 [32] CustomConcept101 that includes several diverse objects such as *plushie bunny*, *flower*, and *chair*. All 101 concepts are utilized.

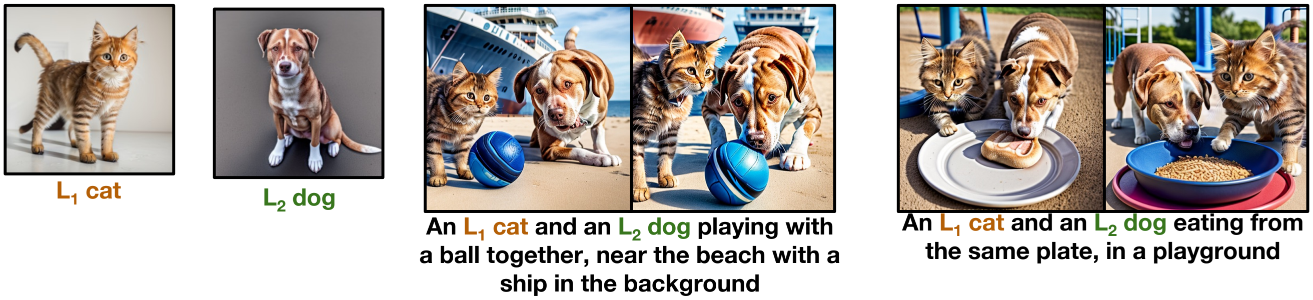
Leveraging the datasets above, we trained LoRAs to represent each concept, totaling to 131 LoRA models. For every competitor, the base stable diffusion model cited in the relevant paper is used. For instance, ZipLoRA [49] employs SDXL, while MixOfShow [16] utilizes EDLoRA alongside SDv1.5. Similarly, our method uses SDv1.5. Note that while the majority of our concepts are derived from CustomConcept101 dataset, the contribution of our benchmark LoRA collection is the 131 LoRA models and additional 200 text prompts.

E.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)

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



L_1 cat
L_2 dog
An L_1 cat and an L_2 dog playing with a ball together, near the beach with a ship in the background
An L_1 cat and an L_2 dog eating from the same plate, in a playground

Figure 11. Qualitative Results showing that CLoRA is capable of composing images in complex interacting scenes.

- 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)

F. Comparison with LoRA-Composer

We compare CLoRA with LoRA-Composer, which operates at test time but requires user-provided bounding boxes, significantly limiting its practicality and ease of use. Additionally, LoRA-Composer is restricted to specific models like ED-LoRA and is incompatible with the wide range of community LoRAs available on platforms like Civit.ai. It also demands substantially more memory, requiring 60GB for generating a composition compared to our method's 25GB for composing two LoRA models. In contrast, CLoRA works seamlessly with any standard LoRA models, including community-sourced ones, without relying on bounding boxes or additional conditions. As shown in Fig. 12, CLoRA consistently produces coherent multi-concept compositions, even in challenging scenarios, ensuring broader compatibility and efficiency. For Fig. 12, the same seed was used for LoRA-Composer with and without bounding boxes to demonstrate the impact of their presence on the results.

G. Multi-LoRA Scalability Analysis

We evaluate our method's performance on compositions with 3-5 LoRA models and compare against state-of-the-art multi-LoRA composition approaches: Mix-of-Show [16] and Orthogonal Adaptation. Both baselines rely on ControlNet key-point conditioning and impose extra overhead: Mix-of-Show trains task-specific Ed-LoRAs, while Orthogonal Adaptation merges all input LoRAs into a single adapter. By contrast, our method is entirely test-time: it needs no key-point cues, no specialized LoRAs, and no merging or retraining.

Table 2 shows quantitative results on 100 compositions with 3-5 input LoRAs. Even without the strong priors such as key-points that explicitly pin down each subject's location, our approach matches the baselines' performance. These results demonstrate that our contrastive test-time strategy maintains high fidelity as the number of subjects and scene complexity increase. Figure 13 shows qualitative results demonstrating CLoRA's capability of generating 2-3-5 people in complex scenes.

		Mix-of-Show	Orthogonal Adaptation	**Ours**
CLIP-I	Max.	0.688 ± 0.042	0.668 ± 0.075	0.668 ± 0.065
	Avg.	0.490 ± 0.031	0.524 ± 0.042	0.525 ± 0.039
	Min.	0.371 ± 0.032	0.395 ± 0.033	0.396 ± 0.034
DINO	Max.	0.574 ± 0.078	0.548 ± 0.093	0.543 ± 0.080
	Avg.	0.351 ± 0.039	0.343 ± 0.066	0.347 ± 0.054
	Min.	0.155 ± 0.046	0.158 ± 0.058	0.161 ± 0.058

Table 2. Quantitative comparison of Mix-of-Show (MoS), Orthogonal Adaptation (Orth) and our method on 3-, 4-, 5-subject generation using CLIP-I and DINO similarity metrics.

H. 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 [60] 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 Tab. 3. The results demonstrate that our method significantly outperforms others across multiple

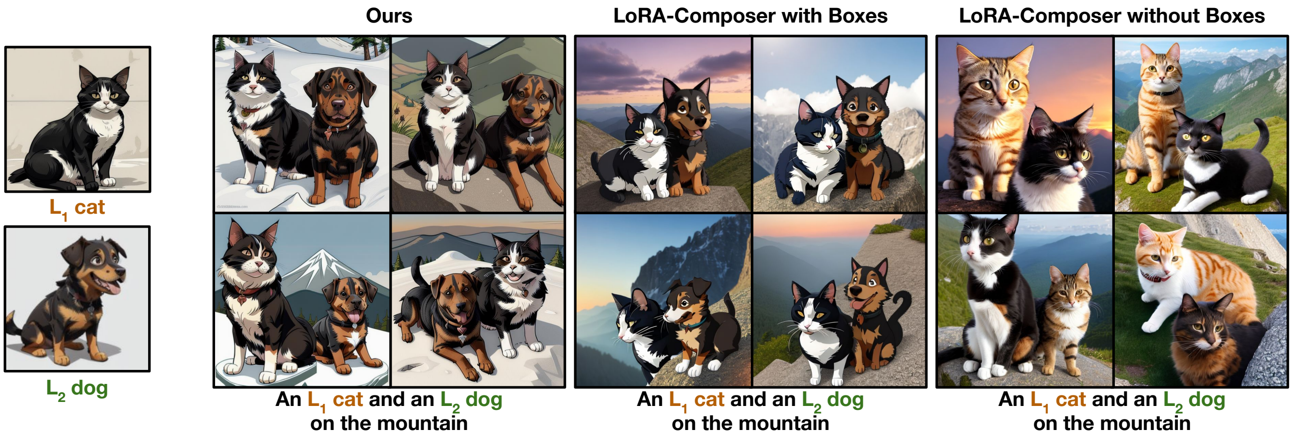

Ours
LoRA-Composer with Boxes
LoRA-Composer without Boxes
L_1 cat
L_2 dog
An L_1 cat and an L_2 dog on the mountain
An L_1 cat and an L_2 dog on the mountain
An L_1 cat and an L_2 dog on the mountain

Figure 12. **Qualitative comparison with LoRA-Composer.** CLoRA achieves consistent multi-concept compositions without bounding boxes, unlike LoRA-Composer. Without user-provided bounding boxes, LoRA-Composer method fails to generate the accurate depictions (see rightmost images).

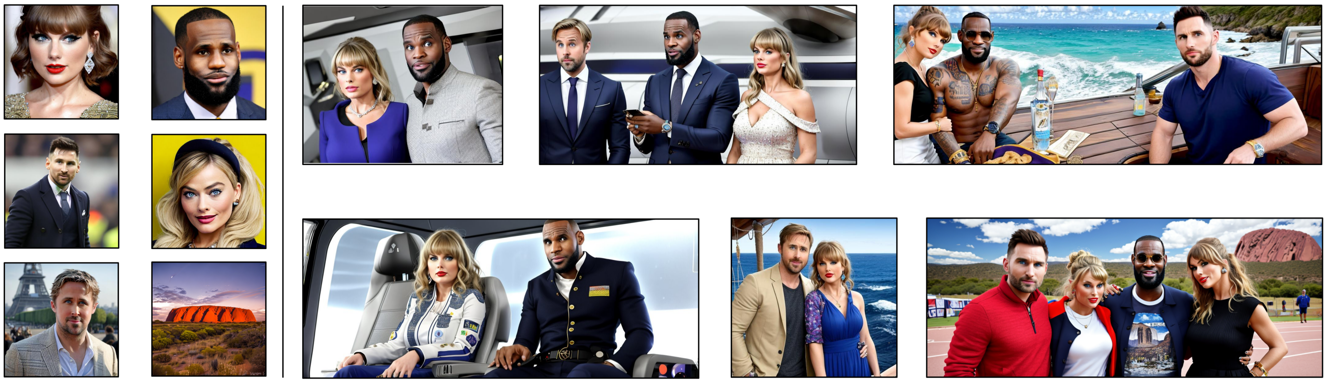


Figure 13. Various compositions using 2-3-5 Lora models and complex scenes demonstrating CLoRA is capable of composing multi-subject images in complex scenes.

Table 3. **Quantitative Comparison of Multi-Concept Compositions.** We evaluate different methods using CLIP and DINO similarity scores on instance segmentation maps. Our method consistently outperforms others across all metrics, achieving the highest minimum, average, and maximum similarity scores.

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

metrics. In particular, we calculate DINO scores, which further 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.

I. Comparison with Orthogonal Adaptation

We compare CLoRA with Orthogonal Adaptation, which operates by enforcing constraints to separate attributes across LoRAs. While this method reduces interference, it requires additional user-provided conditions, such as sketches or key points, to ensure accurate multi-concept compositions. In contrast, CLoRA achieves consistent multi-concept compositions without relying on extra conditions. As shown in Figure 14, Orthogonal Adaptation struggles to generate accurate depictions when such conditions are absent. Our approach seamlessly integrates multiple LoRA modelss while preserving individual attributes, leading to more coherent and natural compositions.

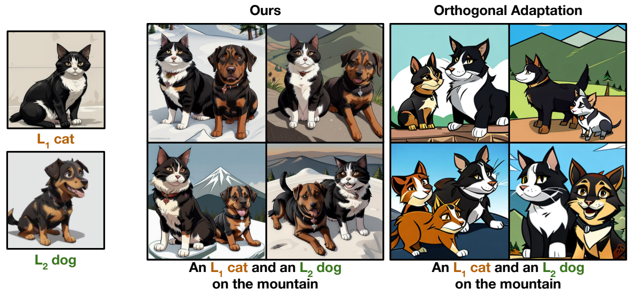

Ours
Orthogonal Adaptation
L_1 cat
L_2 dog
An L_1 cat and an L_2 dog on the mountain
An L_1 cat and an L_2 dog on the mountain

Figure 14. **Qualitative comparison with Orthogonal Adaptation.** CLoRA achieves consistent multi-concept compositions without additional conditions like sketches or key points, unlike Orthogonal Adaptation. Without user-provided conditions, Orthogonal Adaptation method fails to generate the accurate depictions.

J. Additional Qualitative Results

Comparison with OMG. We perform a qualitative comparison between our method, CLoRA, and OMG [30]. OMG relies on off-the-shelf segmentation methods to isolate subjects before generating images. As seen in Fig. 15, 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. 15 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 [5]. 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. 15, 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. 15 demonstrates that CLoRA produces more coherent compositions, maintaining the integrity of each concept even in challenging multi-concept scenarios.

Extensive Qualitative Results. The rest of the Supplementary Materials will provide additional qualitative comparisons which contain the following competitors: Mix-of-Show [16], MultiLoRA [59], LoRA-Merge [47], ZipLoRA [49], and Custom Diffusion [32] on various LoRAs and prompts. Figure 16 compare LoRA-Merge and MultiLoRA using three combined LoRAs, while later figures expand the comparison to include all methods across two separate LoRAs.

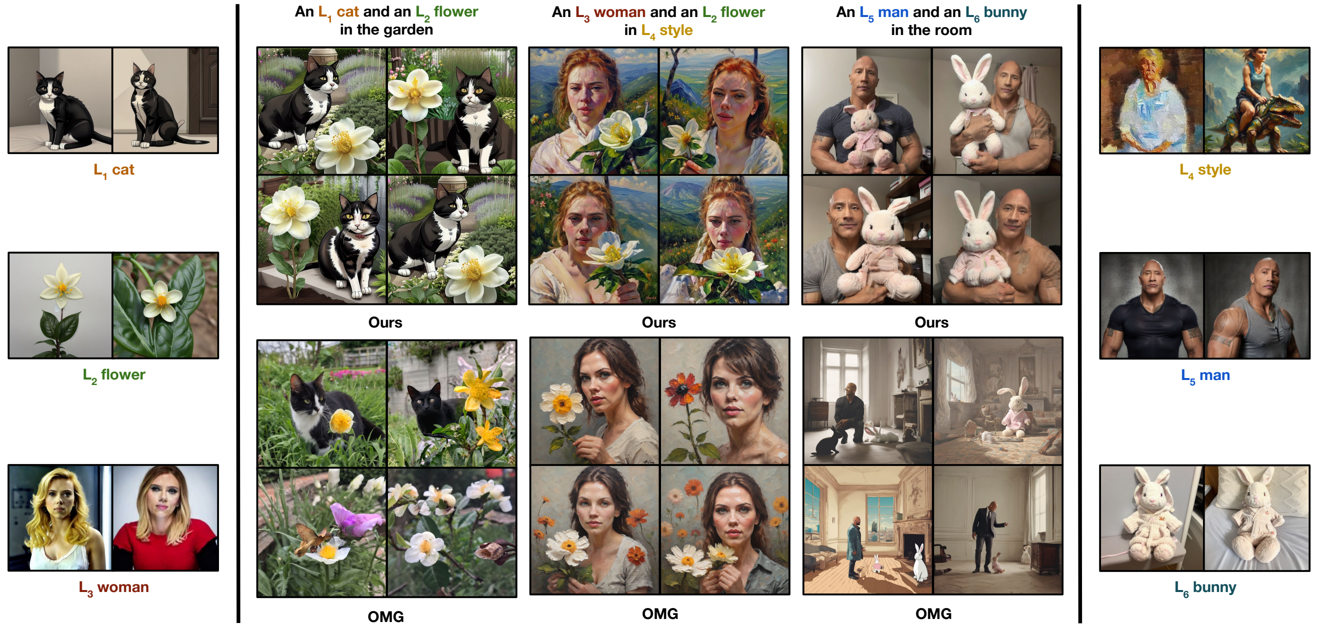

L_1 cat
L_2 flower
L_3 woman
An L_1 cat and an L_2 flower in the garden
An L_3 woman and an L_2 flower in L_4 style
An L_5 man and an L_6 bunny in the room
Ours
Ours
Ours
OMG
OMG
OMG
L_4 style
L_5 man
L_6 bunny

Figure 15. **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.

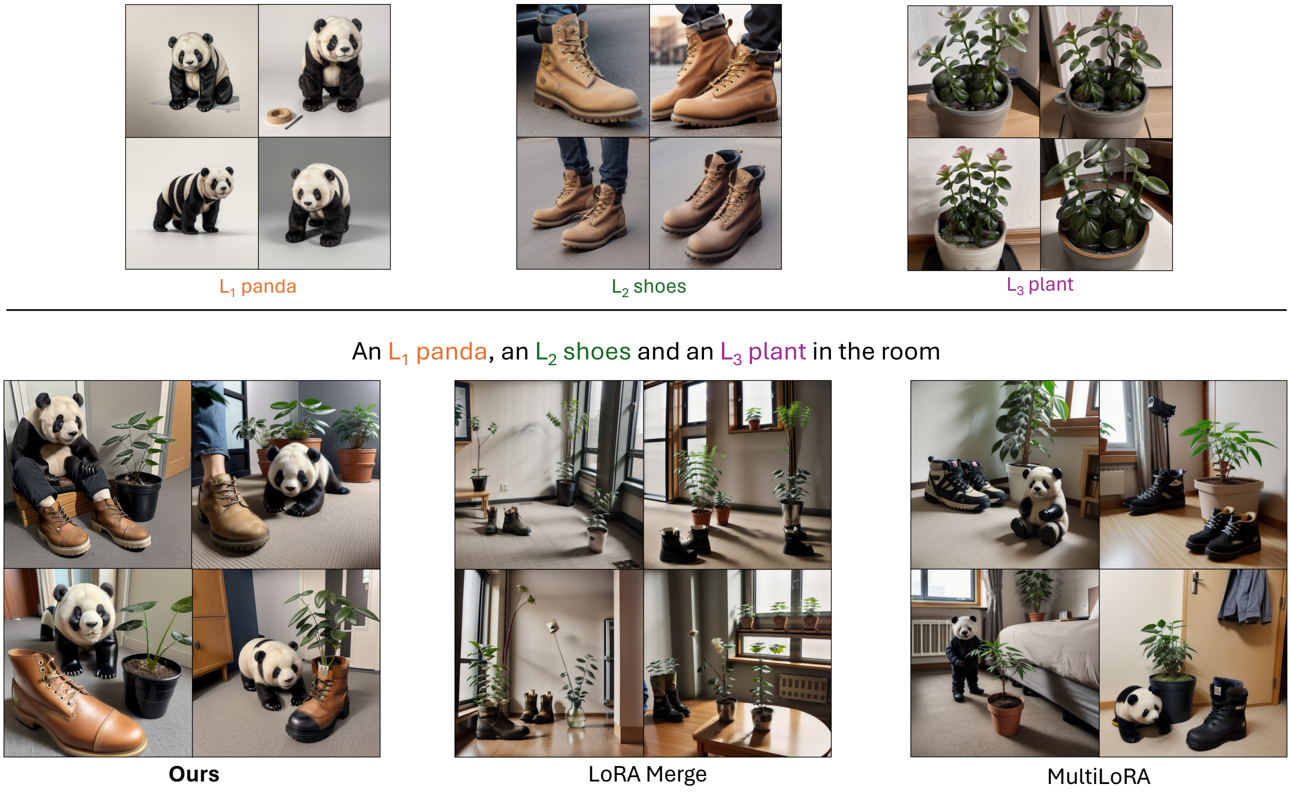

L_1 panda
L_2 shoes
L_3 plant
An L_1 panda, an L_2 shoes and an L_3 plant in the room
Ours
LoRA Merge
MultiLoRA

Figure 16. **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.

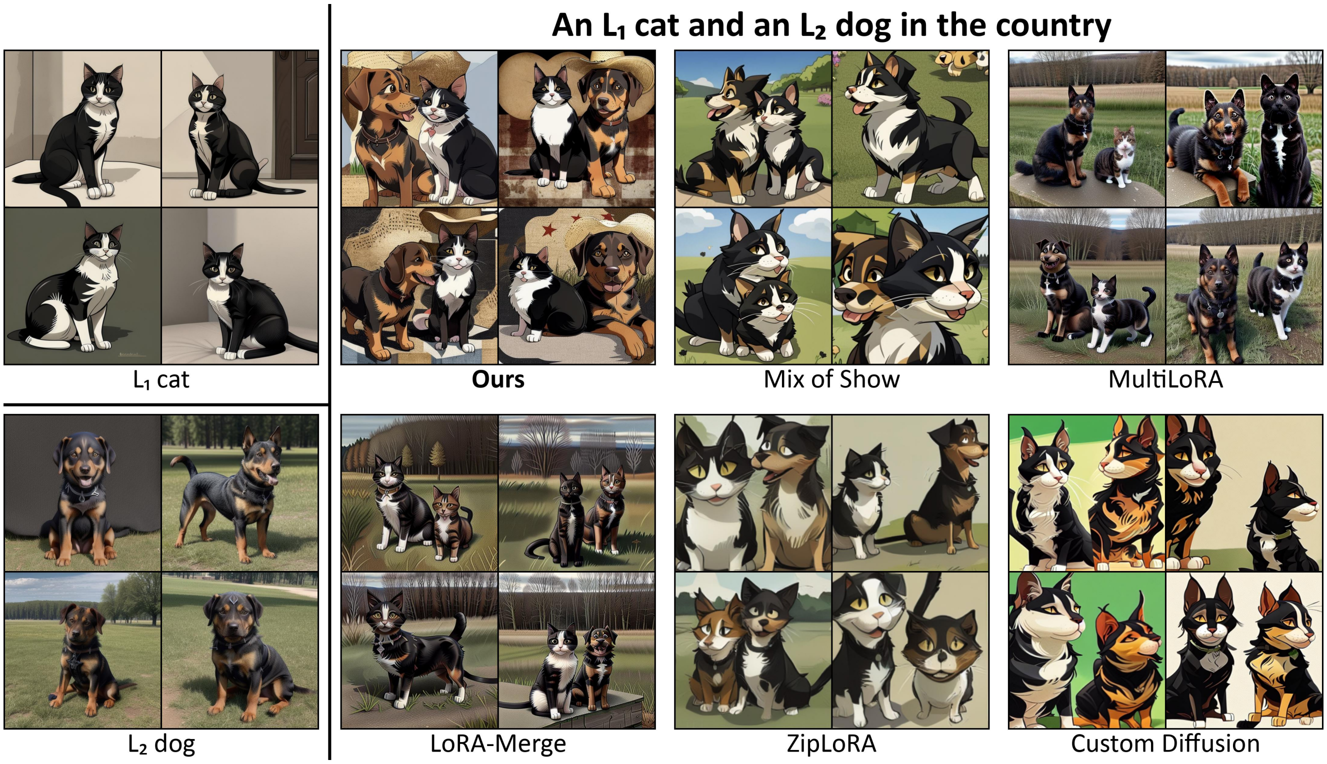


Figure 17. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

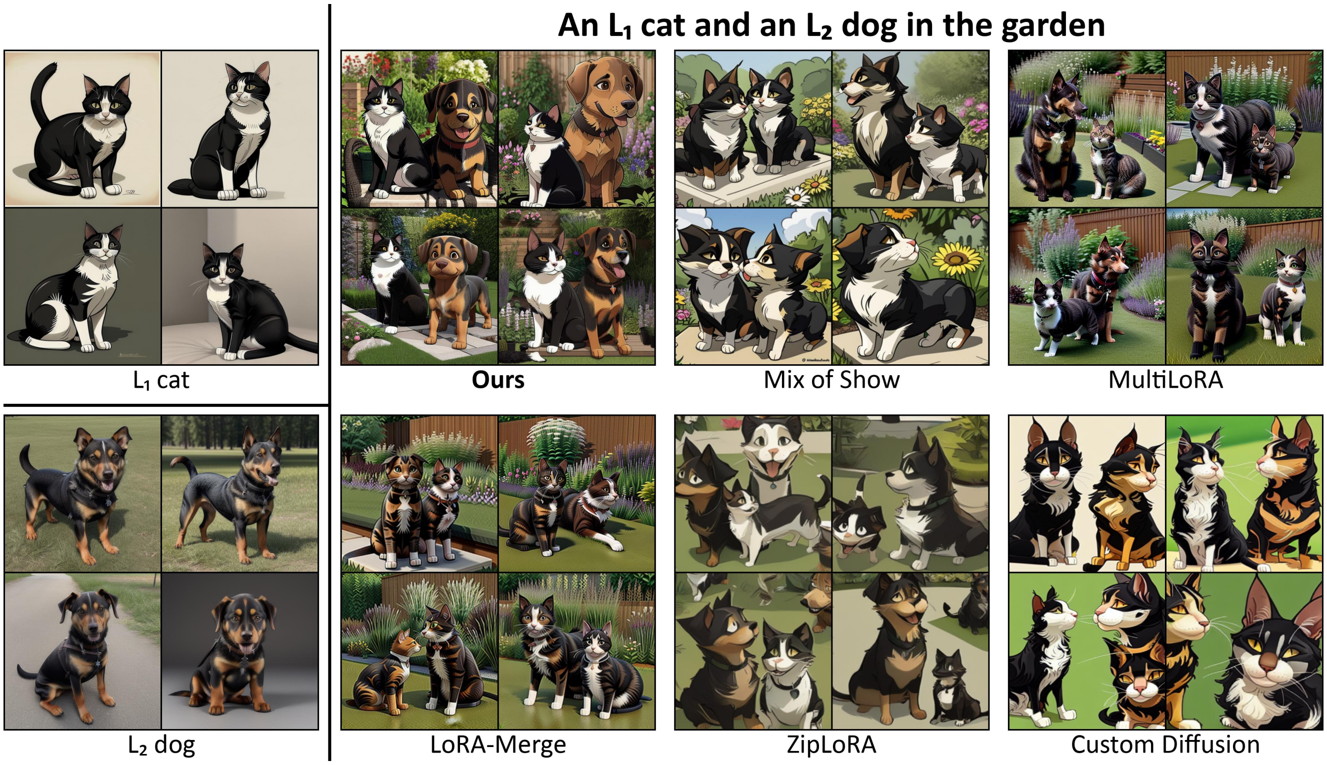


Figure 18. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

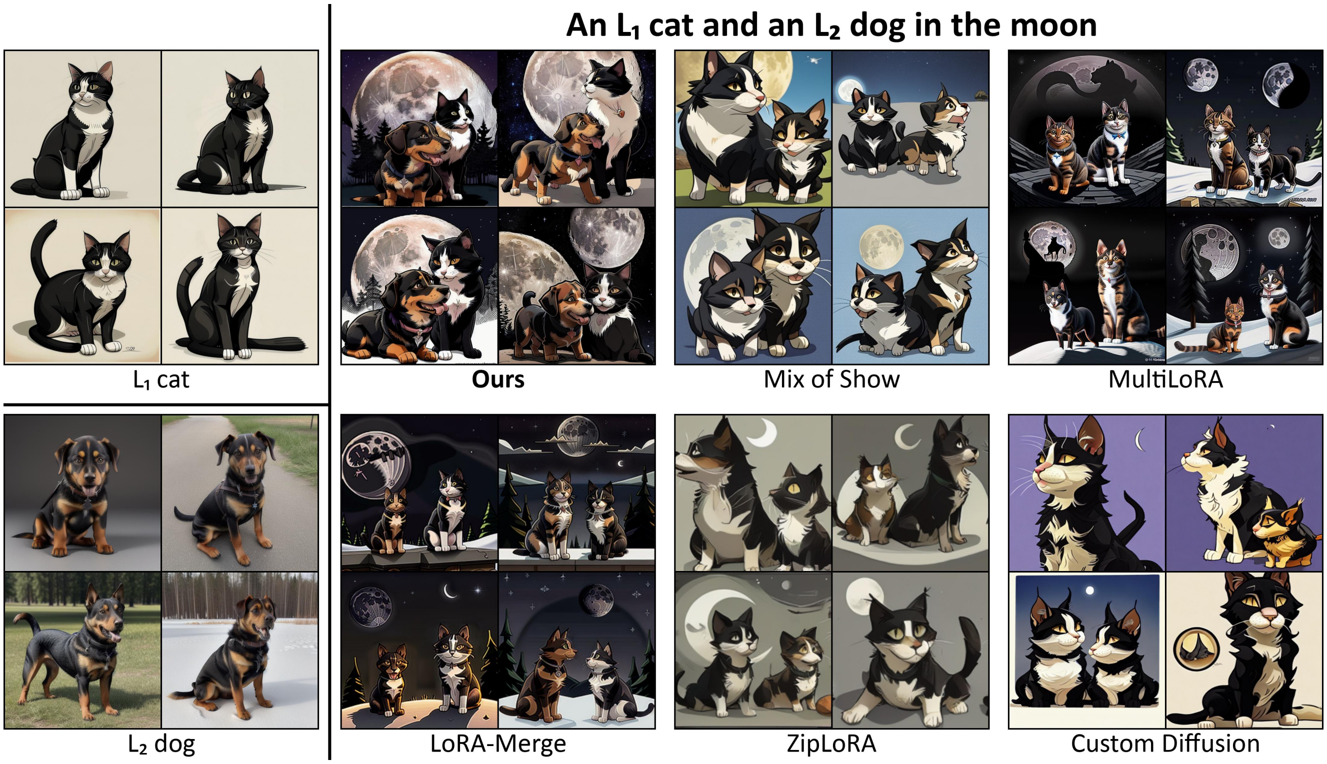


Figure 19. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

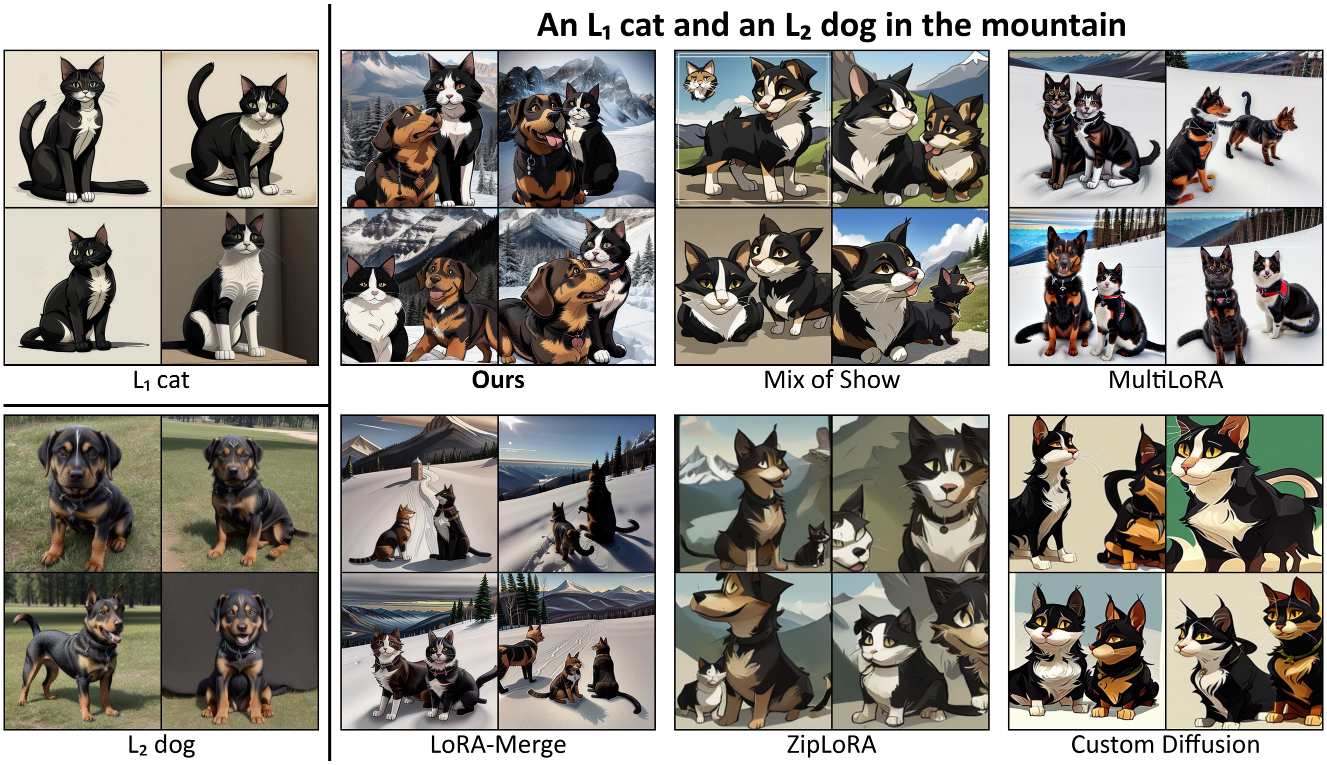


Figure 20. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

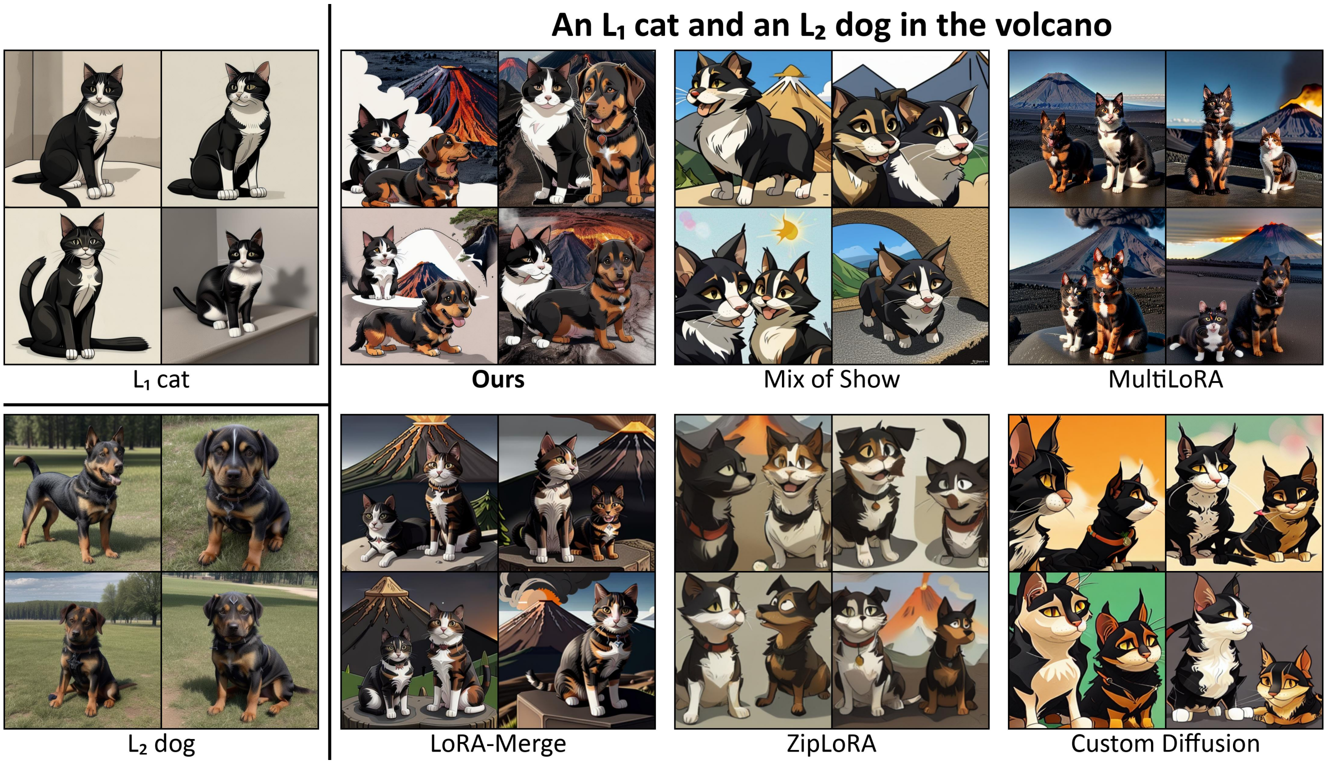


Figure 21. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

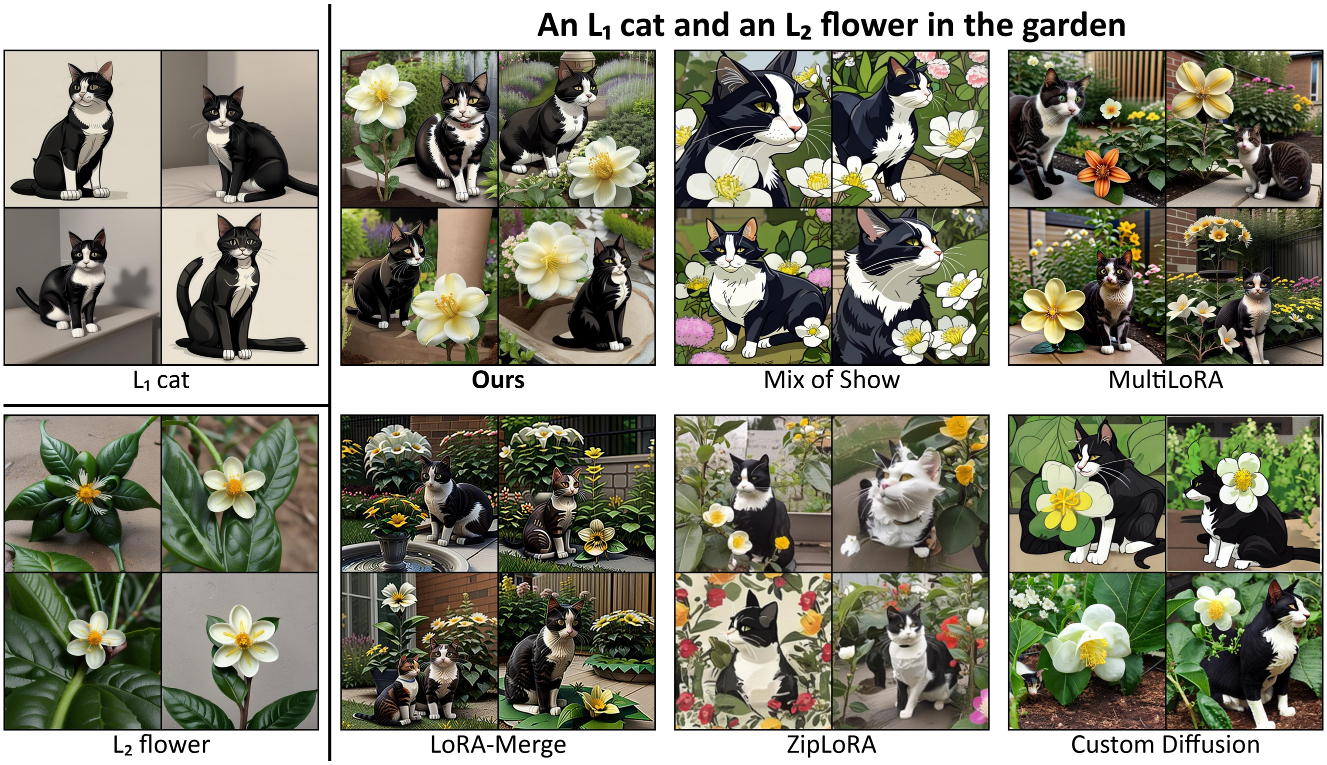


Figure 22. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

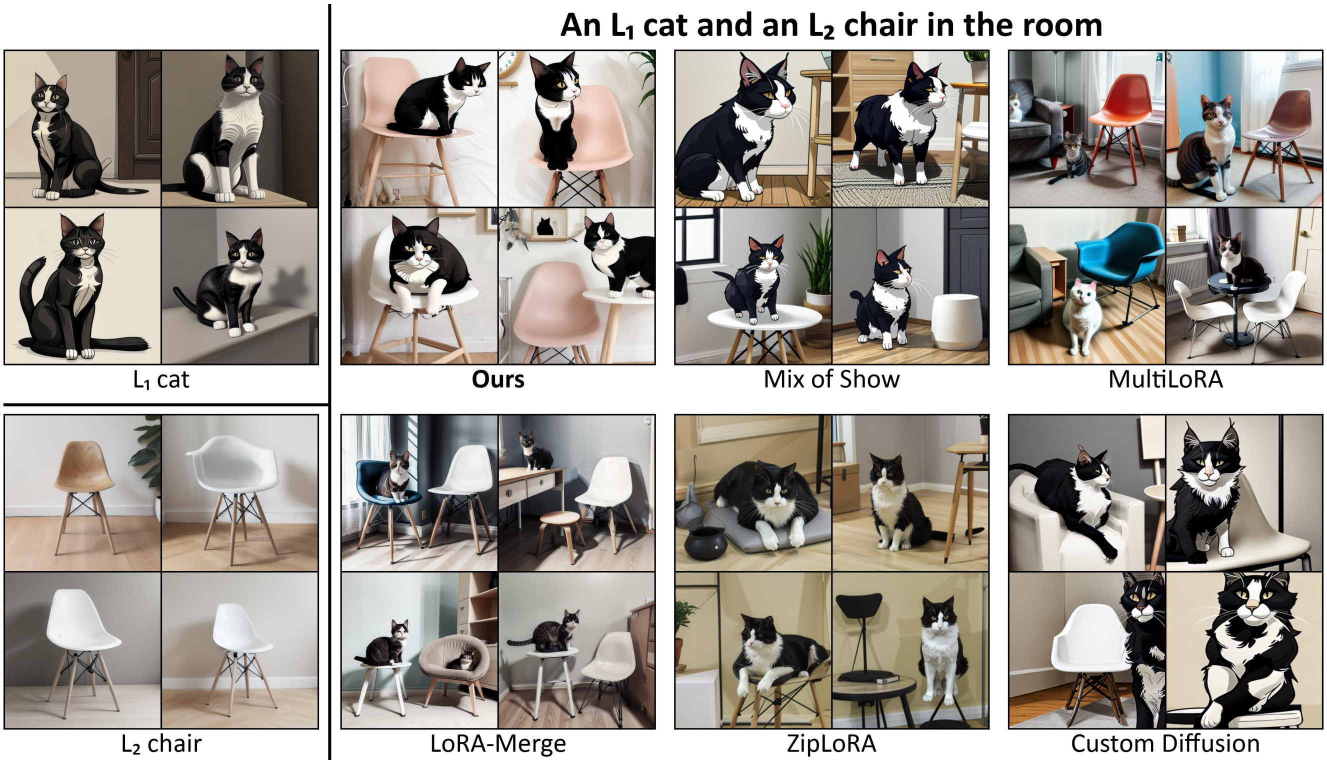


Figure 23. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

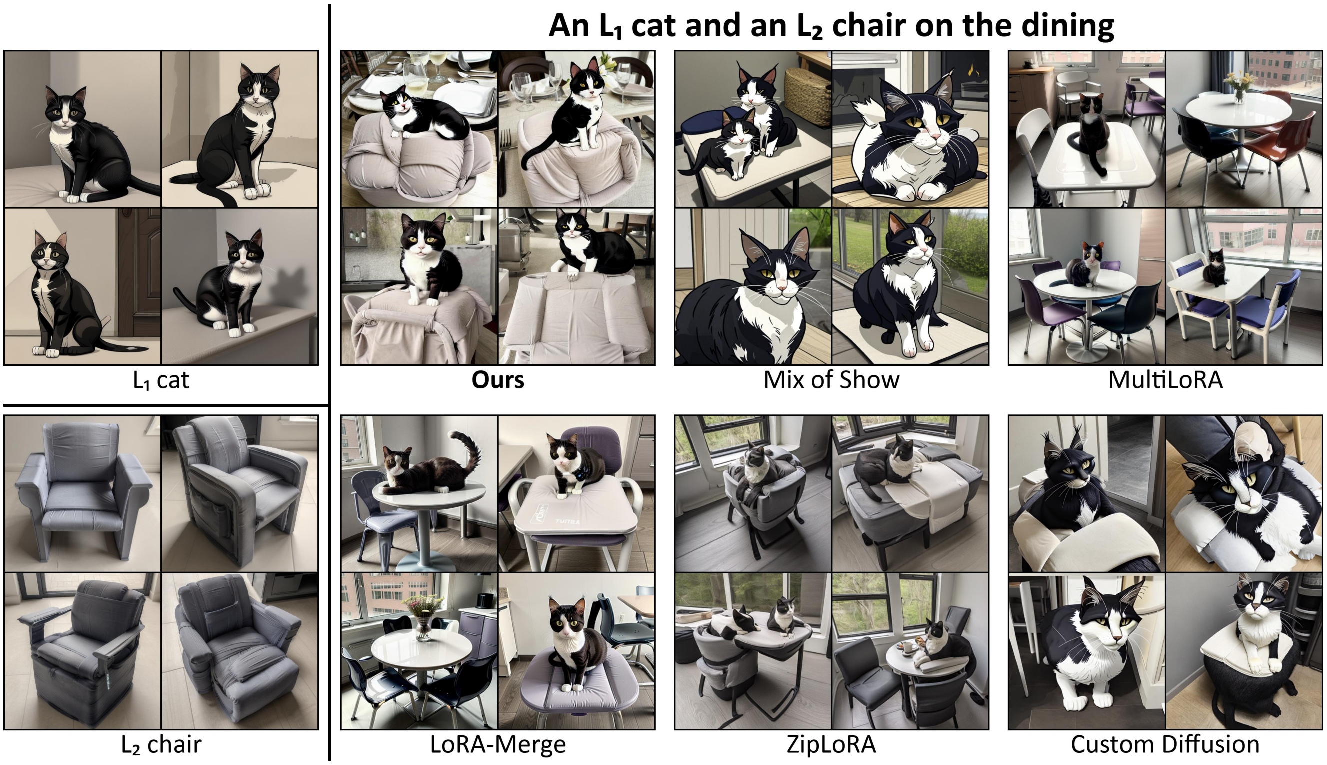


Figure 24. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

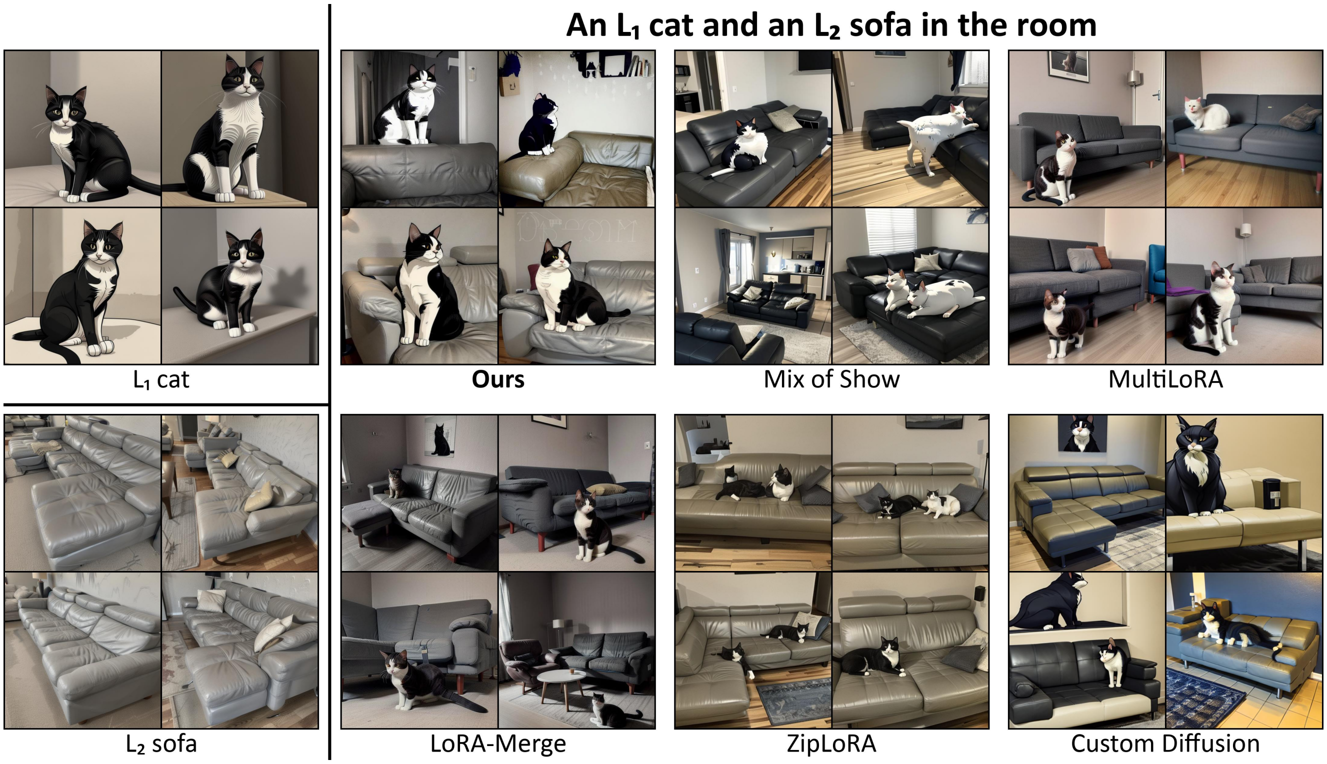


Figure 25. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

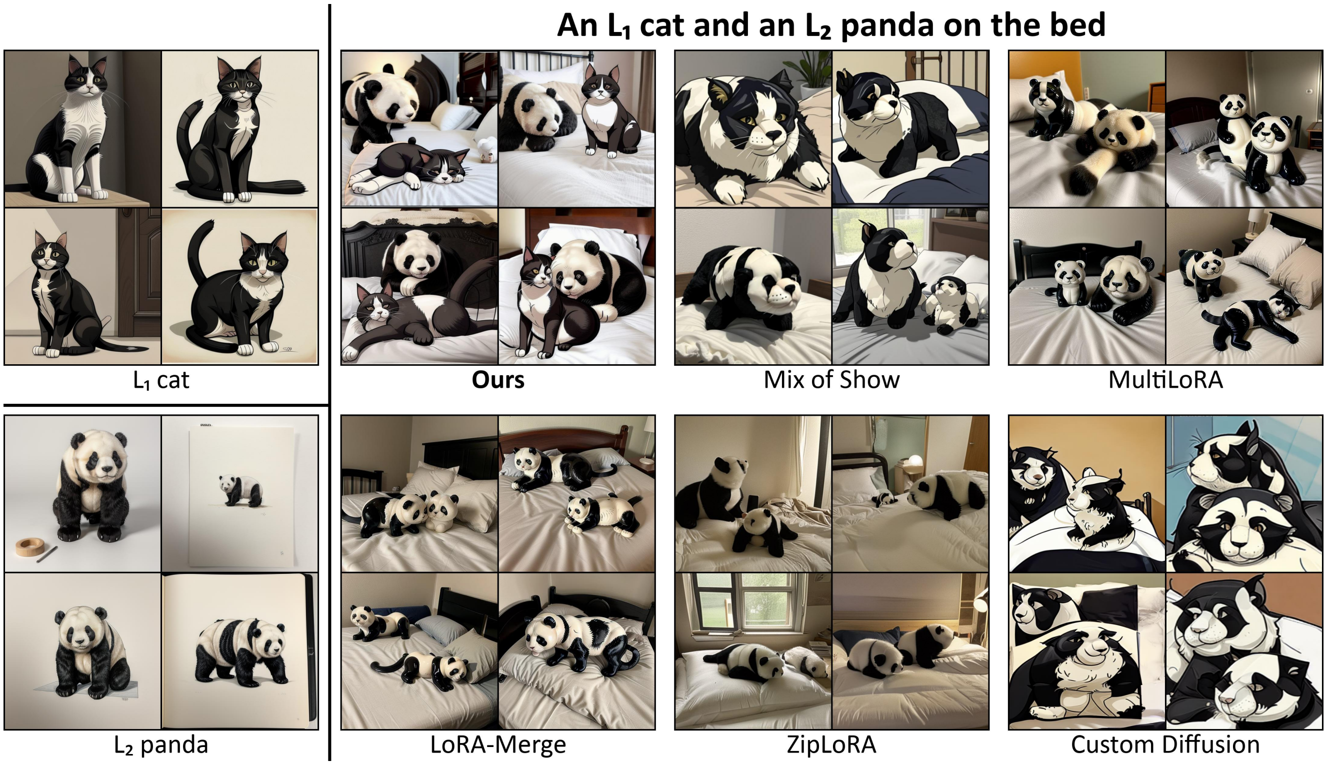


Figure 26. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

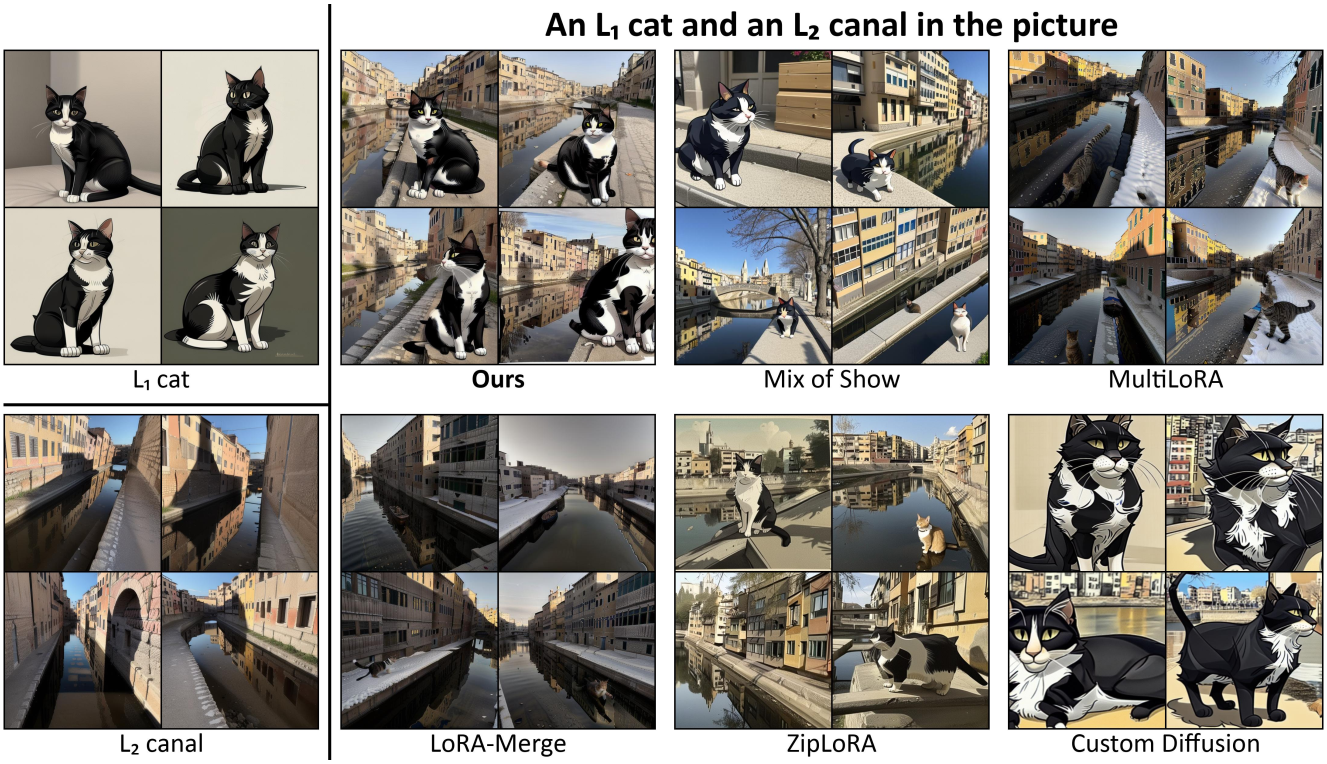


Figure 27. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

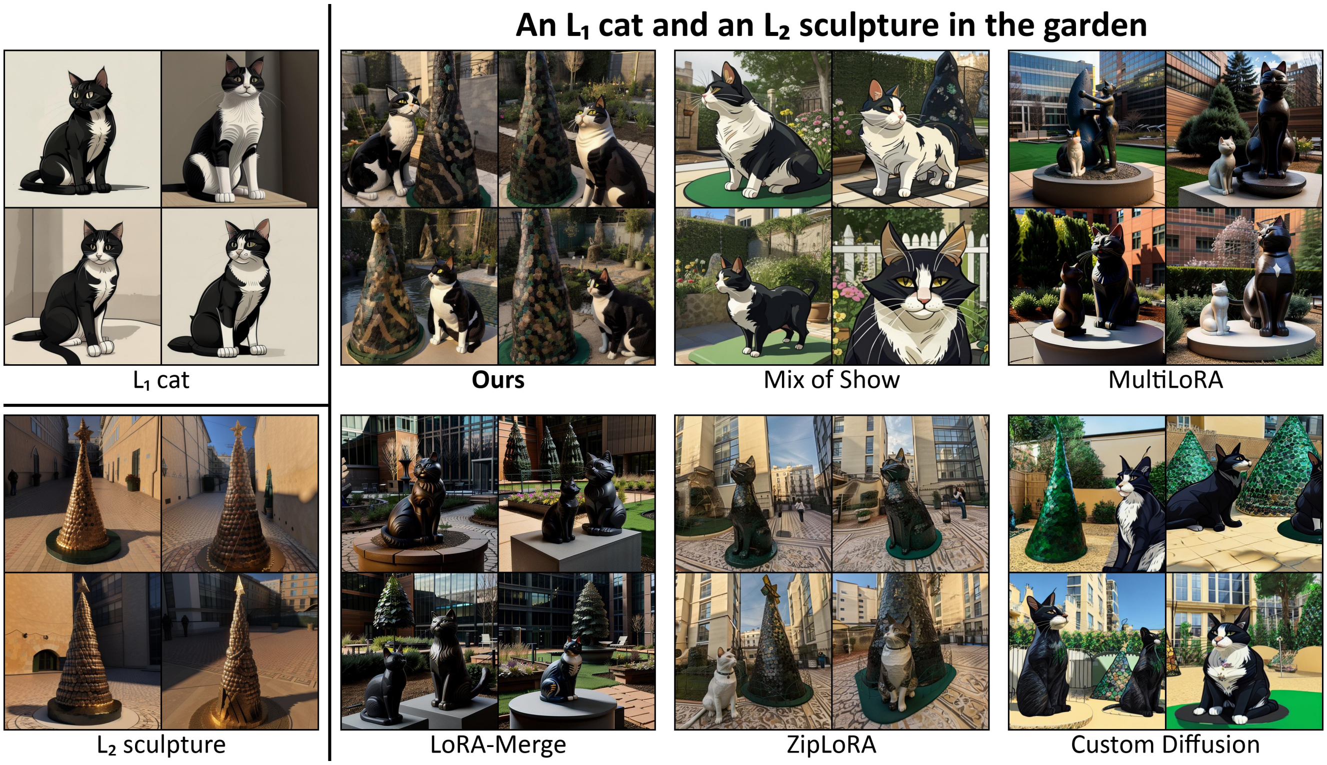


Figure 28. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

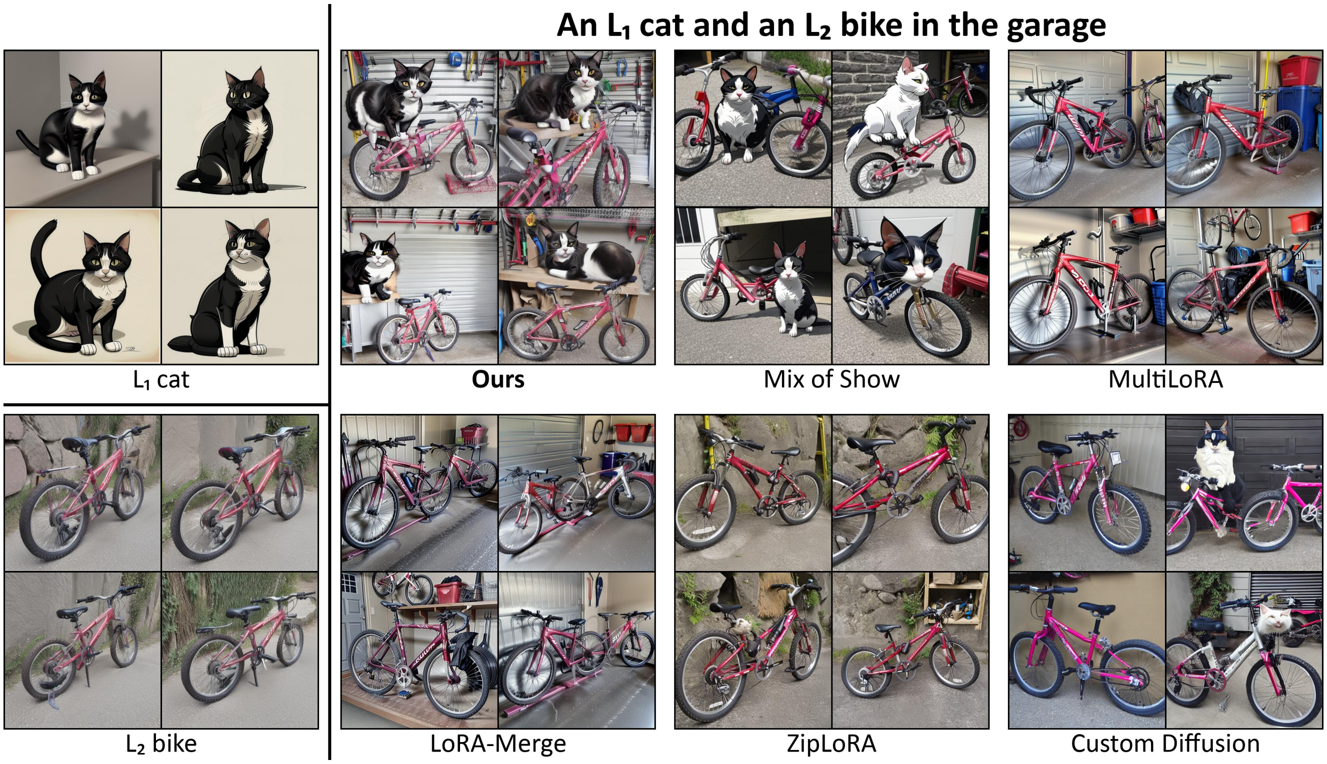


Figure 29. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

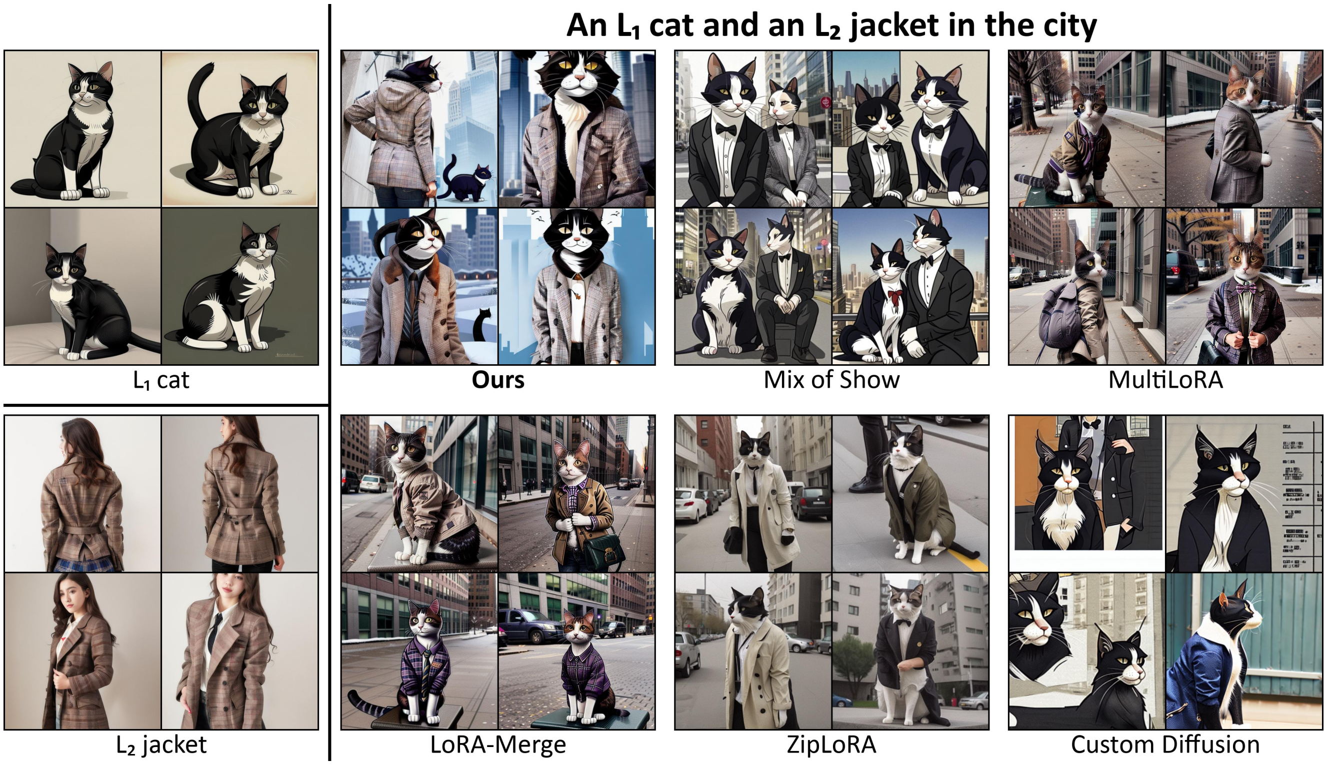


Figure 30. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

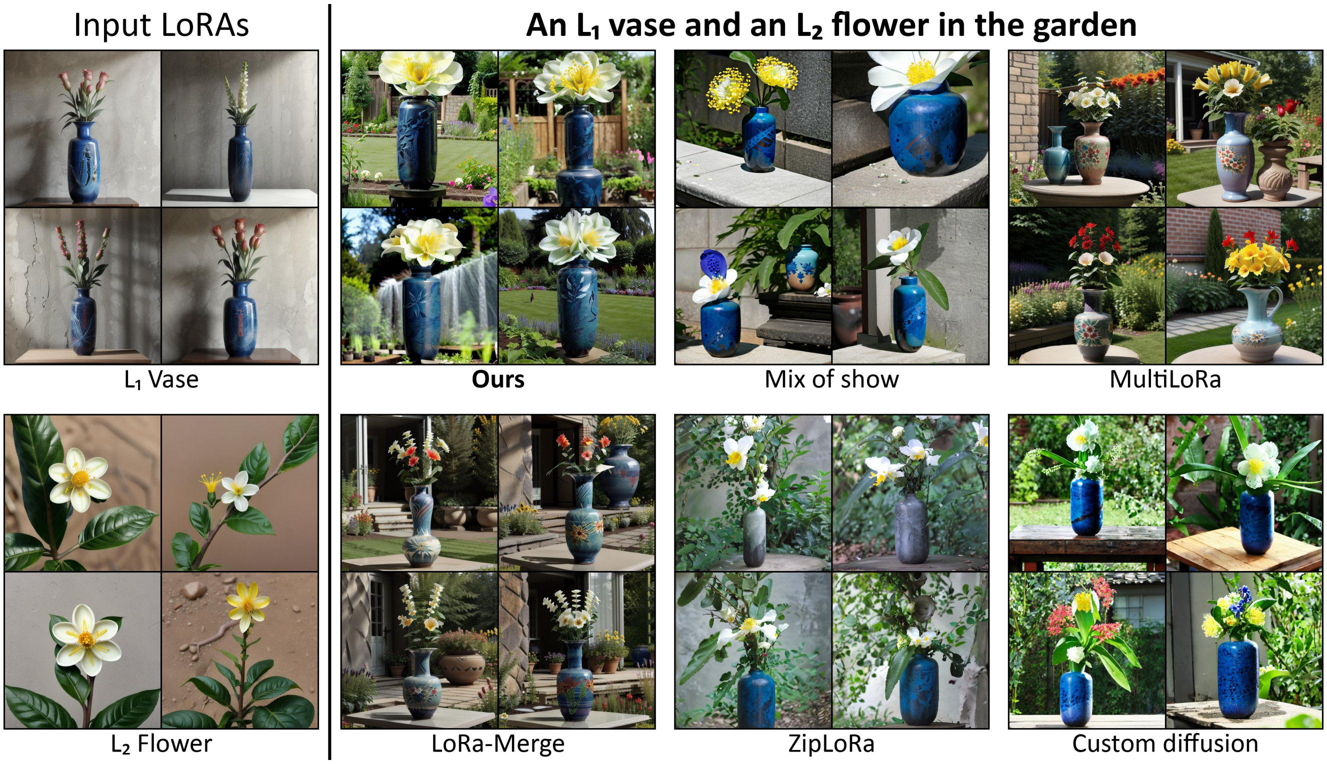


Figure 31. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

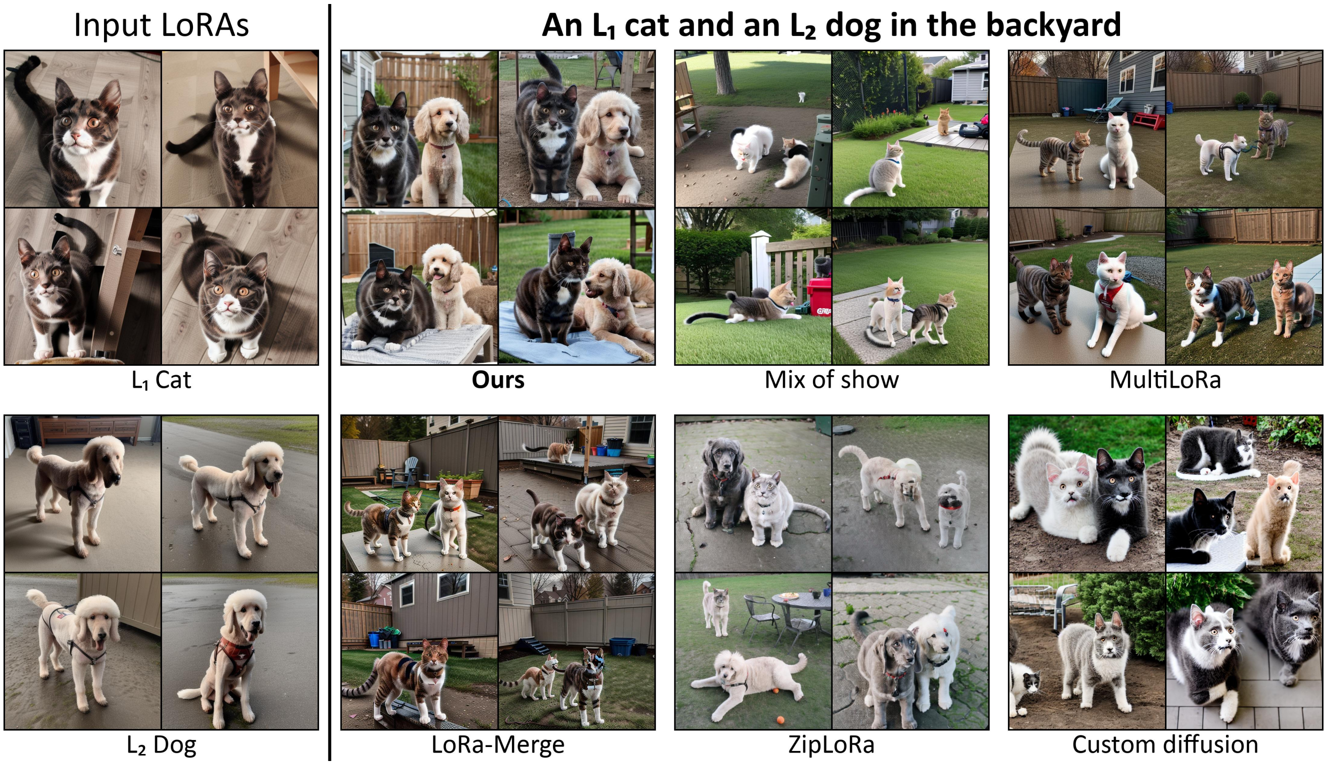


Figure 32. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

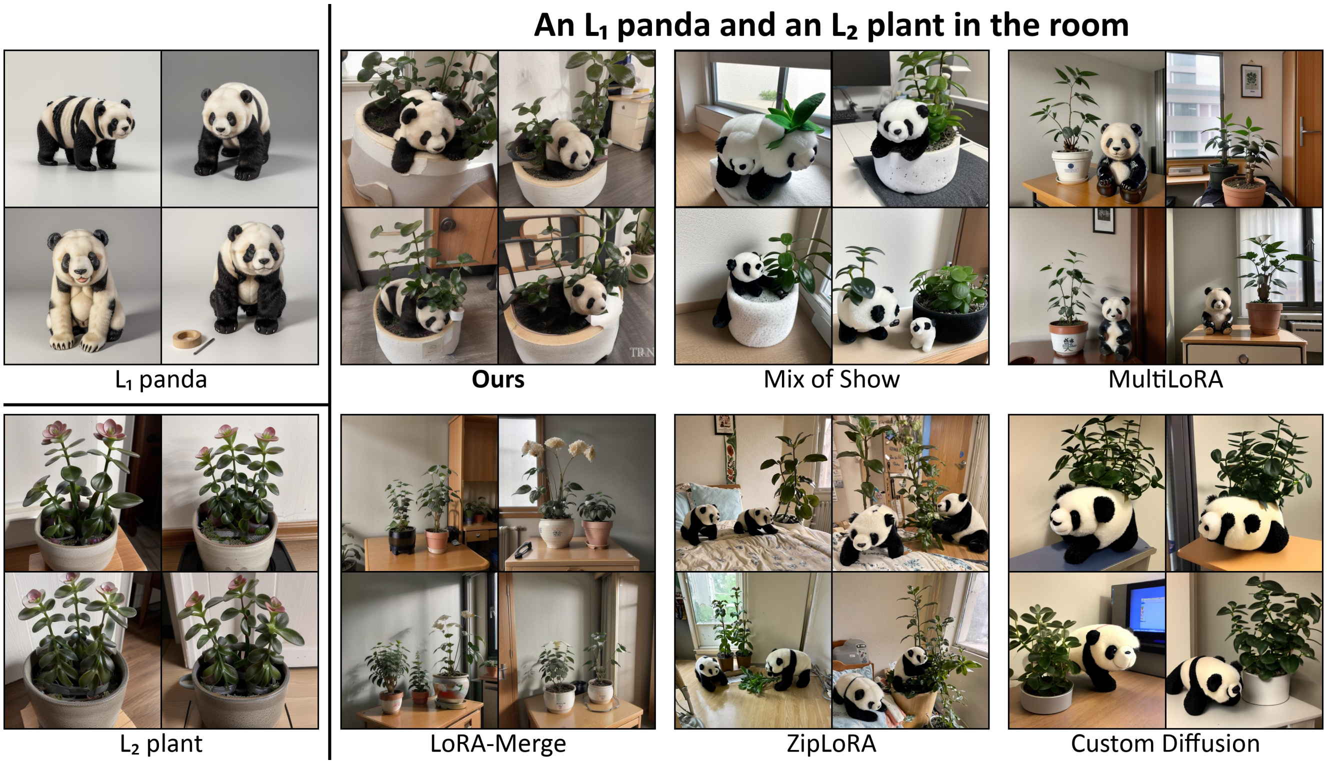


Figure 33. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.

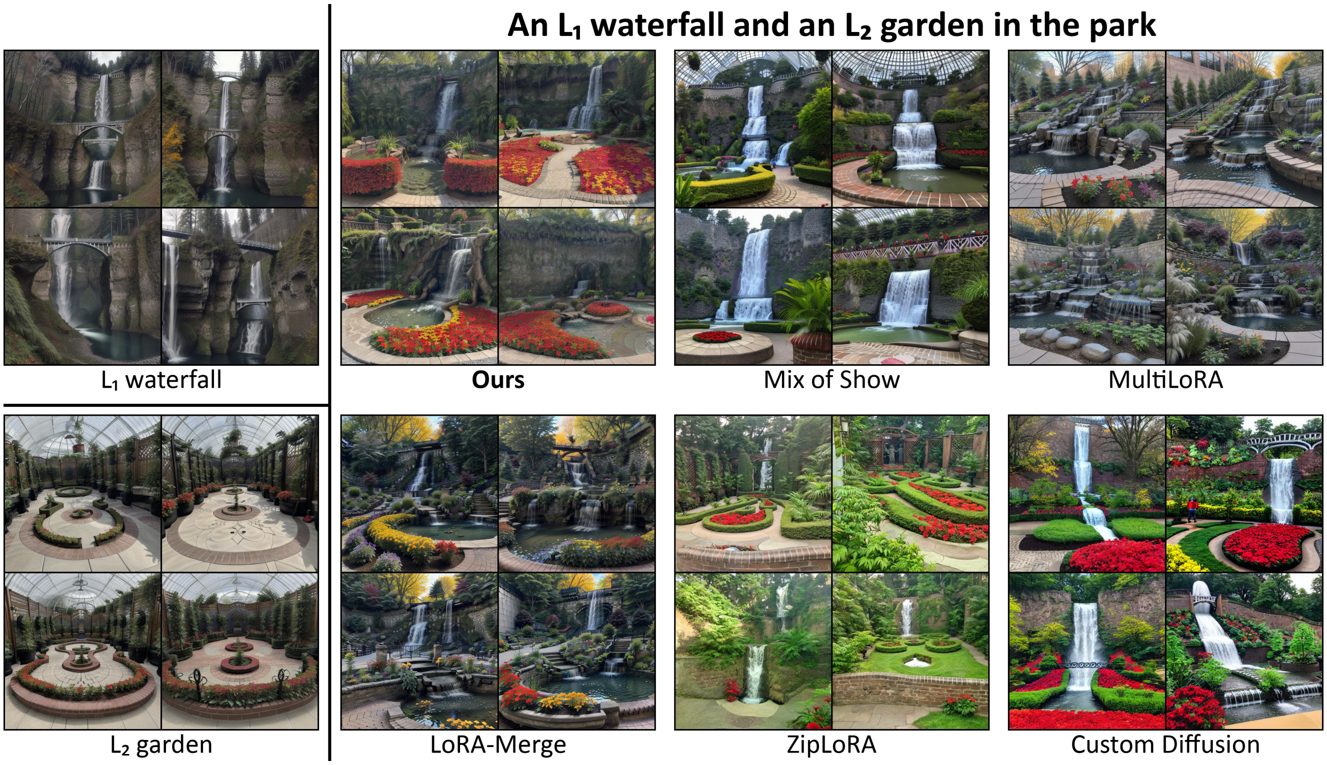


Figure 34. **Qualitative comparison of CLoRA** with other LoRA methods. Our approach consistently produces images that more accurately reflect the input text prompts, LoRA subjects, and LoRA styles.