UnZipLoRA: Separating Content and Style from a Single Image

Supplementary Material

Contents

A. Additional Implementation Details

In this section, we provide additional implementation details for our algorithm:

Block separation strategy. As discussed in Sec. 3.3.3, we employ a block separation technique similar to that proposed in B-LoRA. Specifically, as shown in Fig. 9, the U-Net architecture in SDXL comprises three primary components: the downsampling blocks, the middle blocks, and the upsampling blocks. Each of these components contains several groups of transformer-based blocks. As the upsampling component is more critical to the overall performance, our primary focus for block separation lies in upsampling. Within the upsampling component, there are two distinct groups of blocks, which are differentiated by number of transformers per block: the first group (`Upblock0` in Fig. 9) contains blocks with 10 transformers each, while the second group (`Upblock1`) has blocks with only 2 transformers. Due to this disparity, the first group plays a more significant role in the learning process.

B-LoRA identifies that the first block of `Upblock0` mainly contributes for subject learning, while the second block of `Upblock1` is more specialized for style learning. This observation holds when the input images are relatively simple or lack intricate details. However, when the input images have fine-grained details, only one block is insufficient to capture all the necessary information. To enhance the content learning, we utilize additional block as well. Moreover, we leverage all the blocks in the second group of the U-net upblocks for style learning in order to maintain a balance of the learning capacities between the content and style.

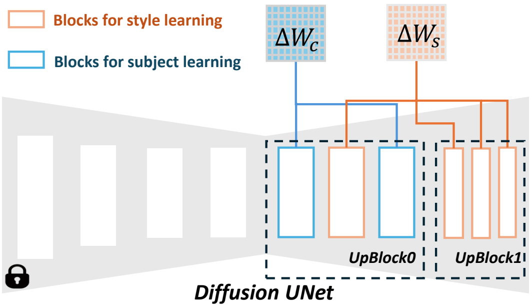

Figure 9. **Block separation strategy in the diffusion U-Net architecture of SDXL.** The illustration highlights the separation of transformer blocks into two distinct groups for disentangling content LoRA ΔW_c and style LoRA ΔW_s.

Reproduction and experimental settings. While reproducing the results of the competing methods, we use the exact hyperparameters reported in their respective papers (or their official implementations), including learning rate, training steps, and other experimental settings. For our approach, most experimental settings — such as the learning rate, batch size, and sampling frequency — are consistent with those described in the main paper. However, the number of training steps required for our method is 600 for most input images, which is lesser as compared to other approaches. Depending on the complexity of the input however, our method may require higher number training steps (in the range of 800 steps) if input image contains fine-grained details.

B. Additional Experiments and Results

B.1. Trigger Phrases Selection

As mentioned in the main paper, we follow the standard prompt construction strategy “a $<c>$ in $<s>$ style” with trigger phrases $<c>$ and $<s>$ in our text prompt to describe the content and style respectively. For selecting the trigger phrases $<c>$ and $<s>$, we follow the descriptor strategies of existing approaches such as DreamBooth [30], StyleDrop [34], ZipLoRA [33], and B-LoRA [6].

As discussed in DreamBooth [30], using a unique token for the subject helps the model to bind the new subject to a unique embedding vector in space since a subject is usually spatially localized in a particular region of the image. On the other hand, style is typically spread throughout the image, thus using just a generic text description increases the flexibility during style training as reported by Style-

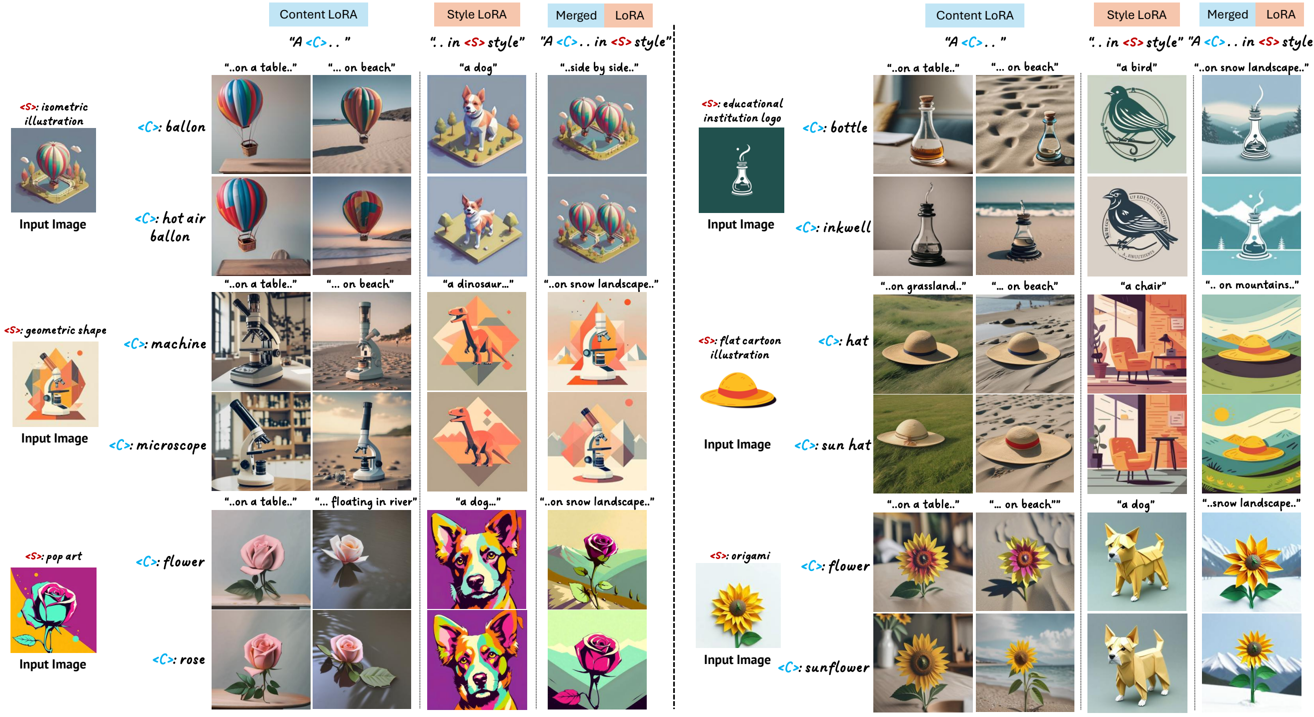

Figure 10. **Trigger phrase selection for the subject**.

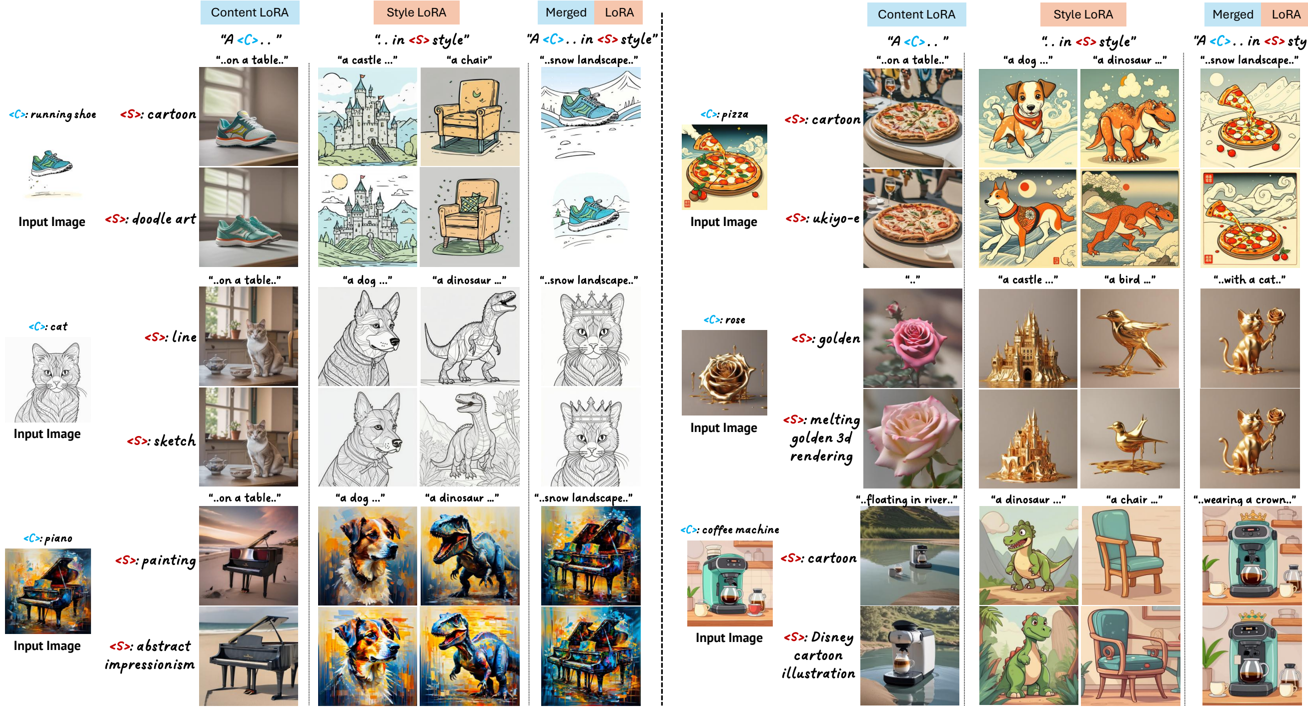

Figure 11. **Trigger phrase selection for the style**.

Drop [34]. This choices are further validated by the follow up works such as ZipLoRA [33] and B-LoRA [6]. Thus, in our experiments we use a unique token followed by the subject class label as the subject trigger phrase (e.g. $<c>$ ='sks

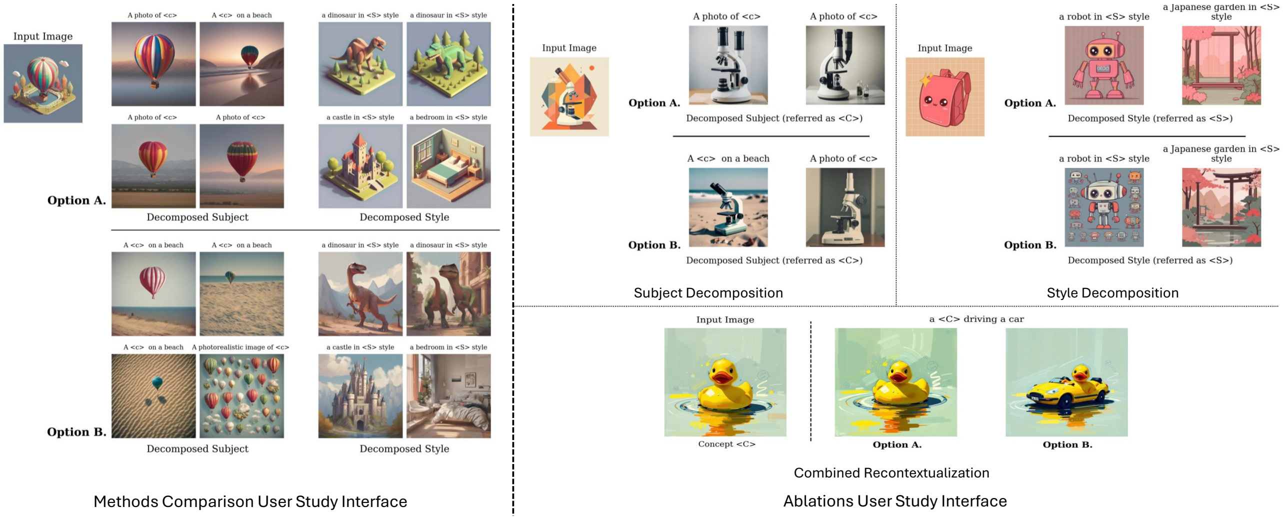

Figure 12. **User Study Interfaces.** We depict the graphical user interfaces (GUI) we used in (left) methods comparison user study for comparing quality of subject-style decomposition, and (right) ablations user study for comparing subject decomposition, style decomposition, and combined recontextualization. User study results are included in the main paper.

dog'), and use generic artistic description for the style (e.g. $<s>$ = 'watercolor painting'). For example, the prompt we use for the first input image in Fig. 13 is "A sks sun hat in flat cartoon illustration style".

Here we independently validate the above choices by studying the impact of the trigger phrases on the results of our method as the degree of detail of these phrases is varied. We confirm that a single-word class label for the subject, and a generic, brief (2-3 word) description for style are sufficient to effectively guide our method, while a more detailed prompt provides additional flexibility for reinforcing any desired attributes.

In Fig. 10 and Fig. 11, we show several groups of examples demonstrating the influence of subject and style trigger phrases on generation. We compare results trained with more general prompts, such as referring to a subject by its category rather than its specific name or using broad and vague style phrases (e.g. using 'flower' instead of 'sunflower'). The results indicate that the generation quality with general prompts is largely preserved, showing no noticeable degradation compared to more detailed prompts.

In some cases, using detailed phrases can help reinforce specific attributes, thus providing flexibility to users. For example, in Fig. 10, using 'flower' as a trigger phrase retains most of the characteristics of the input subject, while using a more detailed prompt 'sunflower' helps boost the fidelity of the features in the center of the flower. Similarly, using 'microscope' instead of generic 'machine' helps retain the fine-grained shape and color characteristics of its eyepiece tube.

Similar conclusions hold for style trigger phrases as well (see Fig. 11): UnZipLoRA can successfully learn the style of the input image even with highly generic, single word style phrases such as 'cartoon' and 'painting'. At the same time, more detailed descriptions can help reinforce specific attributes. While these attributes may still be captured without explicit mention, incorporating them into the prompt ensures more stable preservation. For example, the differences in generations of 'golden' style in Fig. 11 are subtle, yet a more artistically descriptive phrase 'melting golden 3d rendering' leads to clearer stylistic features such as flowing golden droplets. Another example is the pizza in Fig. 11: UnZipLoRA works well with a generic phrase 'cartoon', and specifying 'ukiyo-e' helps retaining more of the background waves, and produces a stronger 'ukiyo-e' aesthetics.

These findings suggest that general trigger phrases suffice for capturing the overall subject and stylistic impression. When retaining specific features is required, explicitly incorporating those features into the input prompt is beneficial.

B.2. Subject-Style Recontextualization Comparison

A key advantage of UnZipLoRA is its ability to produce compatible subject and style LoRAs that can be seamlessly merged via direct addition. This allows for the generation of novel recontextualizations that faithfully incorporate both the subject and style of the original image. Figure 3 in the main paper indemonstrates this capability through various recontextualizations using either individual LoRAs or a combination of both. In this section, we provide qualitative comparisons between our method and B-LoRA for the

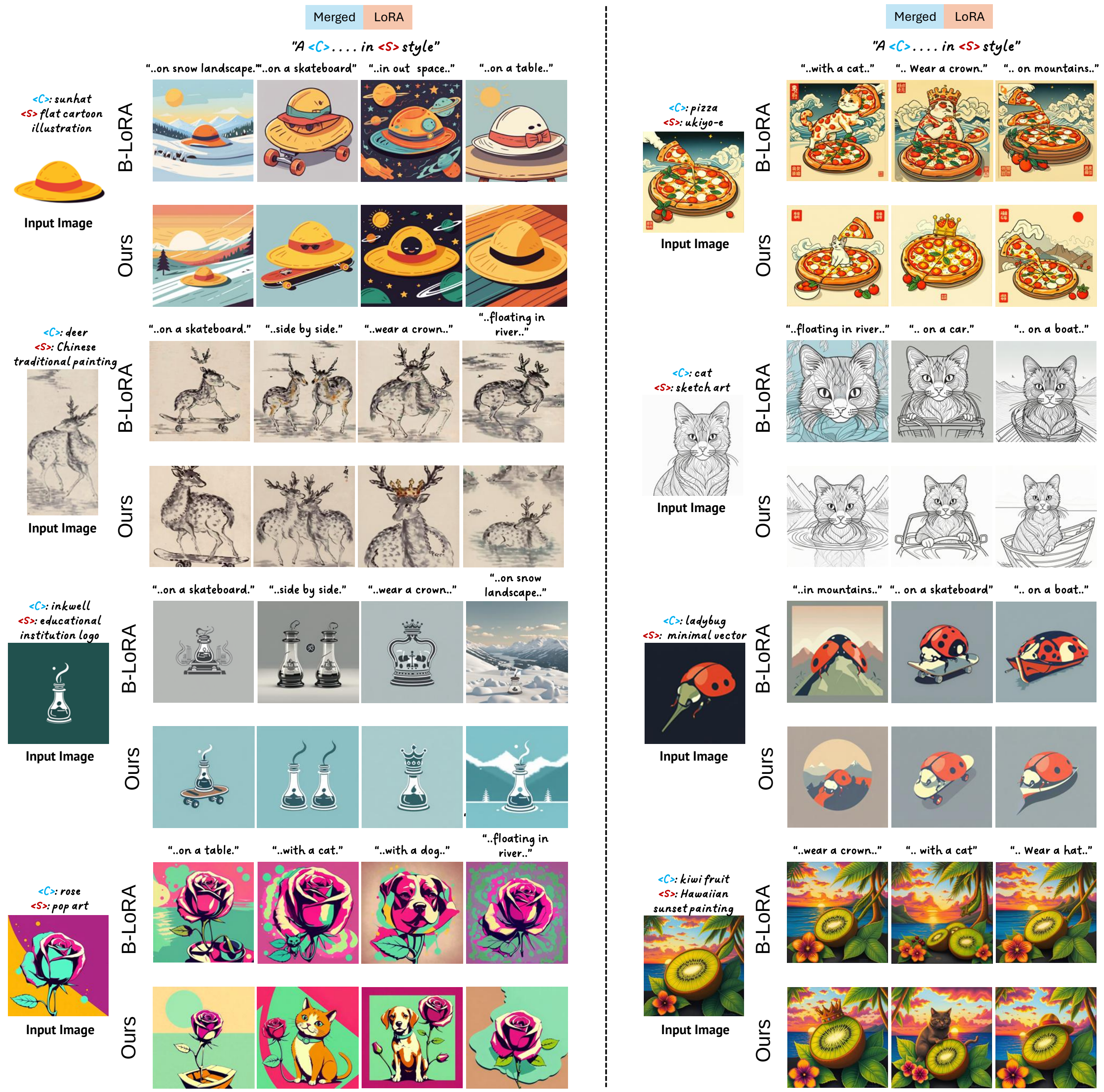

Figure 13. **Comparison of Subject-style Recontextualization**. We present a comparison of subject-style recontextualization between our method and B-LoRA across diverse prompts and input images. The results highlight our method's superior ability to flexibly adapt subjects and styles to various contexts while accurately reproducing both subject and style features.

task of subject-style recontextualization. As demonstrated in Fig. 13, UnZipLoRA is superior in preventing overfitting, reproducing accurate subject and style representations, and enabling flexible recontextualization.

Preventing overfitting. Our method mitigates overfitting by disentangling subject and style representations, ensuring diverse and robust outputs even with challenging prompts

Accurate subject and style reproduction. We achieve precise reproduction of the input's subject and style elements while avoiding blending artifacts.

Flexible recontextualization. Our method enables diverse and logical recontextualization, handling both straightforward prompts like "on a table" and complex, creative prompts that require a nuanced extraction of subject and

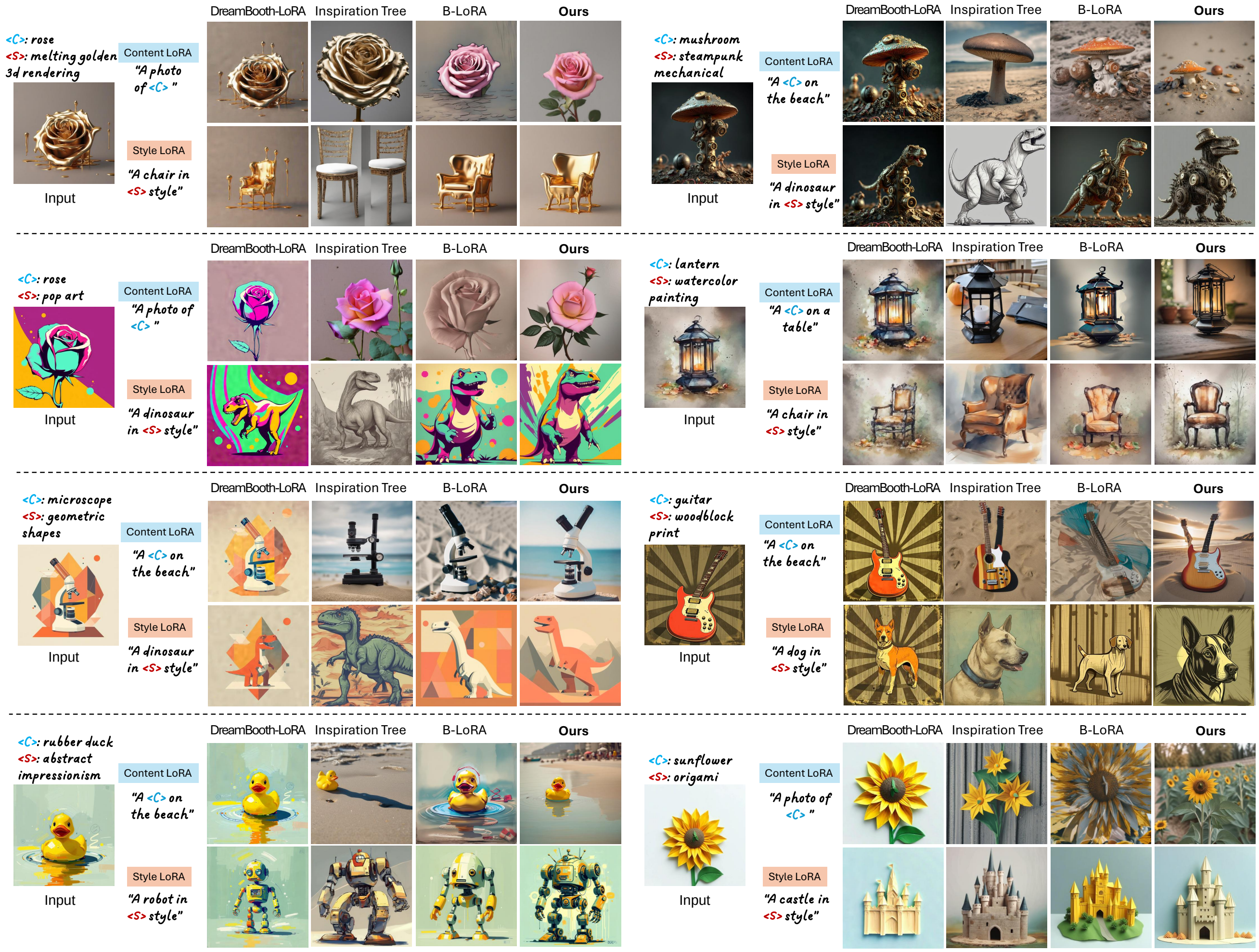

Figure 14. **Qualitative Comparisons.** Additional groups of compare subject and style disentanglement from ours method against DreamBooth-LoRA, Inspiration Tree, and B-LoRA. The result again demonstrates our superior ability to preserve the intended features compared to other methods.

style.

We conducted user studies in the main paper to compare our method with the competing approaches. Beyond the configurations, results, and analyses presented in the main paper, we include the interface used for the main user study and the ablation user study in Fig. 12.

B.3. Additional Qualitative Comparisons

To complement the findings in the main paper, we provide additional qualitative comparisons in Fig. 14 across more diverse prompts and input images. These examples further demonstrate the superior performance of our method in subject-style disentanglement, detail preservation, and overfitting prevention compared to DreamBooth-LoRA, Inspiration Tree, and B-LoRA.

Upon closer inspection of the examples, our observations are consistent with the results presented in the main paper. Each of the compared methods demonstrates limitations that prevent them from achieving the level of disentanglement and flexibility required for our task.

DreamBooth-LoRA. DreamBooth-LoRA struggles to disentangle subject features from style, even though it captures some stylistic features effectively. However, its results often suffer from overfitting to the input image, limiting its ability to recontextualize the style in diverse settings. Our methods successfully captures the style with high fidelity, enabling flexible style recontextualization without overfitting to the input.

Inspiration Tree. While Inspiration Tree effectively prevents the overlap of subject and style concepts and consistently destylizes the input, it struggles to distinguish detailed features of both subject and style. This limitation results in outputs that lack the intricate details of the input subject or style. By incorporating separation strategies, our

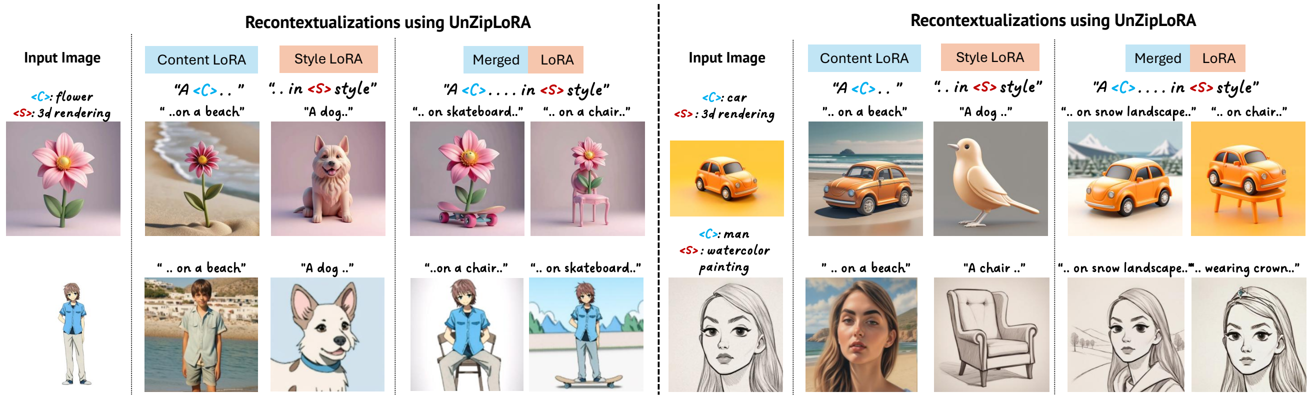

Figure 15. **Additional Decomposition and Re-contextualization Results.** We present additional results on diverse inputs such as humans and 3D rendering styles. The results showcase UnZipLoRA's ability to preserve the details of both the subject and style along with providing flexible recontextualizations.

method intelligently learns and distinguishes these features, leading to more detailed and accurate outputs.

B-LoRA. As what we discussed in Experiments section of main paper, and in Sec. B.2, B-LoRA fails to generate consistent results and suffers from overfitting to input images. While it can always accurately learn style features, it struggles to reproduce subject details reliably. For instance, in the guitar and sunflower examples, B-LoRA fails to consistently retain the original input's color.

Moreover, it often mixes subject and style features, resulting in generations that incorporate unintended elements, such as the background color of the microscope or the sunflower's color being treated as part of the style. In contrast, our method addresses these issues with carefully designed separation techniques, ensuring consistent, disentangled outputs that faithfully represent both subject and style.

B.4. Additional Qualitative Results on Diverse Inputs

We present additional decomposition and recontextualization results using a wider variety of subjects/styles, including more complex subjects such as humans and styles such as 3D rendering in Fig. 15. The results demonstrate that our method generalizes effectively to detailed and diverse subjects, maintaining the fidelity of the subjects/styles, and flexibly recontextualizing them across various contexts, showcasing the superiority of our method in handling rich and challenging inputs.

B.5. Additional Ablation Study Results

We provide additional examples of ablation study of the effects of adding each component, as shown in Fig. 16. Each row demonstrates the recontextualization of subject, style and subject-style across different configurations and baseline methods.

Prompt separation. Distinct prompts and LoRA weights for subject and styles guide their respective LoRAs, ensuring effective and disentangled subject-style decomposition.

Column separation. Dynamic column masks selectively activate relevant columns during training, preserving learning capacity with fewer columns and preventing interference. Enhanced orthogonality between LoRAs improves flexibility in recombination.

Block separation. Style-sensitive blocks effectively capture essential stylistic features, while subject-sensitive blocks focus on fine details.

Table 4. **Ablation Study Alignment Scores.** Comparisons for Content and Style Decomposition among each separation strategy.

	DB-LoRA	M1	M2	M3 (UnZipLoRA)
Style-align. (CLIP-I) ↑	0.417	0.409	0.407	**0.427**
Subject-align. (DINO) ↑	0.339	0.346	0.347	**0.349**
Style-align. (CSD) ↑	0.245	0.217	0.216	**0.265**
Subject-align. (CSD) ↑	0.338	0.352	0.354	**0.358**

Quantitative results. Beyond the ablation user study results provided in Tab. 3 in the main paper, we provide additional quantitative results in the form of subject- and style-alignment scores in Tab. 4. The results confirm each separation strategy's independent contribution: prompt separation (M1) aids subject learning, column separation (M2) improves disentanglement, and block separation (M3) significantly enhances stylization. These trends align with the user study results (Tab. 3) and our analysis in Sec. 4.4 of the main paper.

User study interface. We conducted user studies in the main paper to compare our method with the competing approaches. Beyond the configurations, results, and analyses presented in the main paper, we include the interface used for the main user study and the ablation user study in Fig. 12.

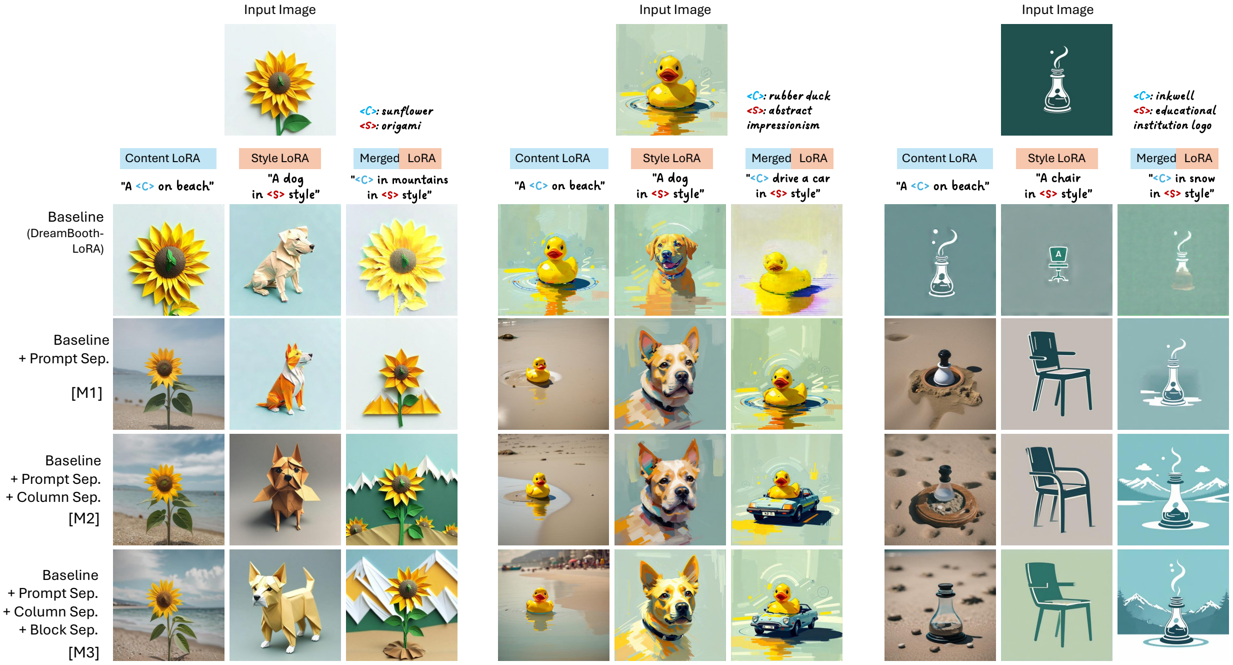

Figure 16. **Ablation study.** We show the impact of adding prompt-wise separation, column-wise separation, and block-wise separation sequentially. Each row illustrates the outputs of baseline methods and our proposed approaches, highlighting their difference in subject-style disentanglement, style fidelity, and recontextualization across different examples.

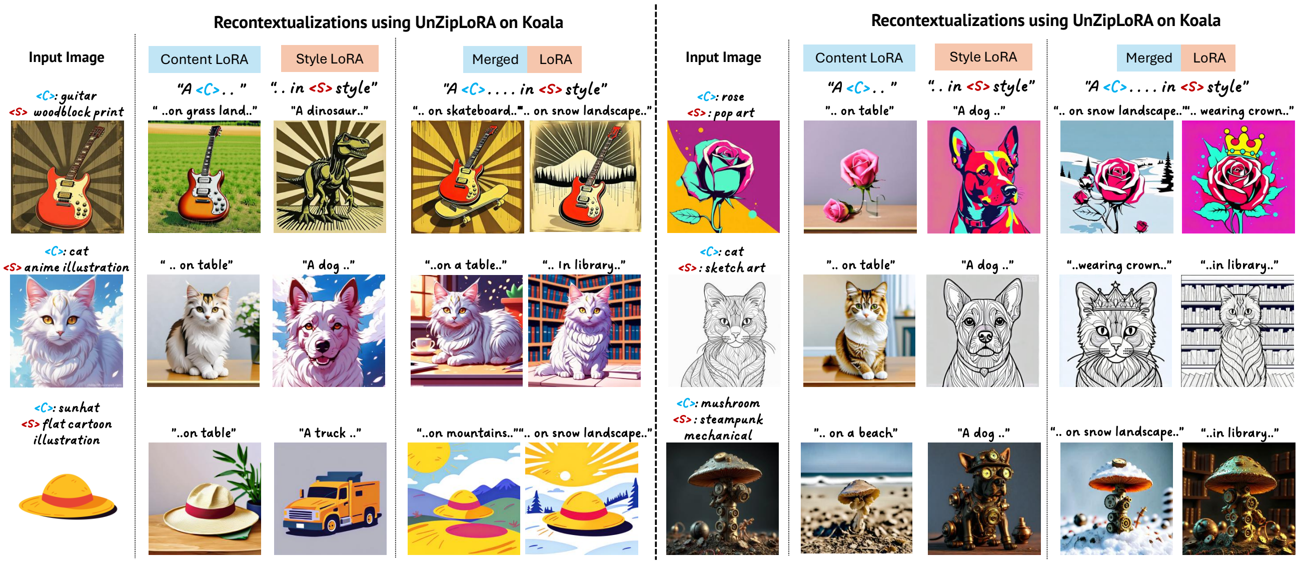

Figure 17. **Additional Results on KOALA Diffusion.** Our approach generalizes effectively, as demonstrated by successful results on the more recent KOALA diffusion model.

B.6. Ablation on Percentage of High-importance Columns

To evaluate the influence of the column selection percentage (N) within our dynamic importance recalibration strategy, we conducted ablation studies across varying values of N. This strategy selects the top $N\%$ of most important columns by calculating the column importance from the gradient information using the Cone method [22]. The results of our ablation experiments shown in Fig. 18 confirm that a selection percentage within the range of $20 \leq N \leq 40$ is sufficient for successfully capturing both subject and style characteristics.

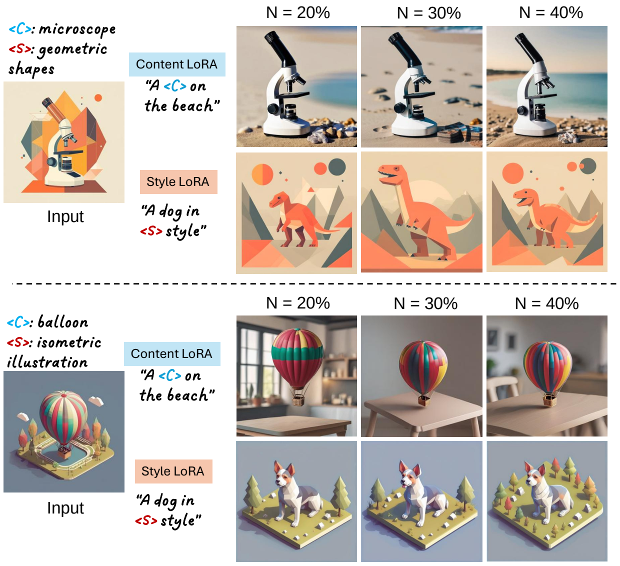

Figure 18. Ablations on different values of N. Our column selection strategy selects top $N\%$ of most important columns usiing dynamic importance recalibration strategy. As shown, $20 \le N \le 40$ is sufficient for capturing subject/style successfully.

B.7. Additional Cross-combination Results

The subject and style LoRAs produced by UnZipLoRA open up a possibility for cross-combination: pairing a subject LoRA from one image with a style LoRA from another. Fig. 19 provides additional results for such cross-combination where the LoRAs are combined by direct addition. While these LoRAs are not explicitly trained together (and thus not subject to the orthogonality constraints enforced by ZipLoRA [33]), the inherent separation imposed by our column and block strategies generally results in higher compatibility than generic DreamBooth-LoRAs trained without such constraints. Consequently, direct arithmetic merger yields promising cross-stylization results.

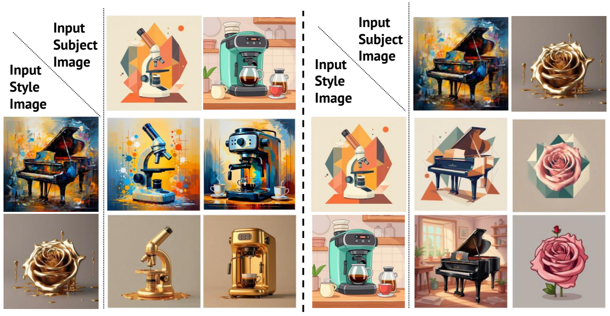

Figure 19. Additional examples of *cross-composition* using subject and style from different input images. The result demonstrates that our method effectively integrates features from both subject and style.

B.8. Additional Results on KOALA Diffusion

Our method extends beyond SDXL, and is applicable to newer diffusion models. We demonstrate this generalizability by training on KOALA [19], a more efficient, recent text-to-image model with leaner architecture compared to SDXL. We provide additional results in Fig. 17. As shown, our method, when applied to KOALA, accurately captures subject and style and allows for successful recontextualization (though the overall quality of the results is not as high as for SDXL due to limited capacity and lower parameter count of KOALA).

C. Compute Requirements

UnZipLoRA achieves strong computational efficiency through a joint training strategy that optimizes resource utilization. By training both subject and style LoRAs concurrently in a single run, UnZipLoRA significantly reduces the overall training time. Specifically, UnZipLoRA requires only 1260 seconds to train both LoRAs on a single NVIDIA A40 GPU.

In contrast, most existing methods necessitate separate training processes for content and style, effectively doubling the time requirements. For instance, DreamBooth-LoRA requires 1860 seconds per LoRA, resulting in a total training time of 3720 seconds. While B-LoRA demonstrates faster individual LoRA training at 600 seconds per LoRA (1200 seconds total), UnZipLoRA remains highly competitive. Notably, methods like Inspiration Tree incur significantly higher computational costs, requiring 7680 seconds in total: 3840s to select a good random seed for training and another 3840s to train the model.

Beyond time efficiency, UnZipLoRA minimizes the number of parameters updated during training. Through its block and column separation strategies, UnZipLoRA updates only up to 30% of parameters in the downsampling block and bottleneck, and approximately 50% in the upsampling block for each LoRA. This focused optimization reduces the trainable parameters by nearly 40% compared to training two full LoRAs independently, further contributing to its efficiency. Owing to such efficient parameter utilization, UnZipLoRA exhibits faster convergence, requiring only 600 steps of training as opposed to 800 to 1000 steps for most other methods including DreamBooth-LoRA and B-LoRA.

D. Image Attributions

In our experiments, we use several stylized images as inputs images. We curate these input images from three sources: (i) free-to-use online repositories that provide artistic images; (ii) open-sourced repositories of previous works such as StyleDrop [34] and RB-Modulation [29]; and (iii) synthetically generated images using freely available text-to-image models such as Flux. For the human-created artistic images, we provide image attributions below for each image that we used in our experiments.

Image attributions for the stylized images used as inputs

The sources of the style images that we used in our experiments are as follows:

- Sun hat in flat cartoon illustration style,
- Kangaroo in one line art illustration style,
- Backpack in cartoon illustration style,
- A bear in kid crayon drawing style,
- A teapot in mossaic art style,
- A telephone in line drawing style

All the remaining input images are generated using Flux[2] text-to-image diffusion model using the prompts provided by RB-modulation codebase on their github page at this URL: Text-prompts to generate stylized images

[2] https://huggingface.co/black-forest-labs/FLUX.1-dev