

# LoRACLR: Contrastive Adaptation for Customization of Diffusion Models

## Supplementary Material

### 1. Evaluation for Multi-concept Generation

Although we follow the same evaluation setup as other competitors, we acknowledge that instance-based metrics may better capture accuracy and identity for multi-concept image generation. To further supplement our current metrics, we develop a pipeline to quantify composition accuracy and identity. Our pipeline begins by segmenting individual subjects using the model from Li et al. [3], allowing us to isolate key elements in both generated and ground truth images. Once segmented, each subject is assigned to a corresponding concept based on extracted feature similarities, ensuring a meaningful comparison between generated outputs and reference subjects. We then compute three instance-based metrics to evaluate the composition. Accuracy measures correct subject reconstruction by comparing segmentation masks, while Identity quantifies how well the generated subject preserves its identity using DINO-based feature similarity, which captures fine-grained semantic consistency. Additionally, CLIP-I Similarity assesses alignment between generated subjects and their ground truth counterparts using CLIP embeddings. LoRACLR demonstrates significant improvements across these metrics, reinforcing our qualitative findings, see Tab. 1.

Table 1. **Quantitative comparison of instance-based metrics.** Our approach achieves higher scores across all metrics, indicating better subject disentanglement and identity preservation.

	Method	Average	Min	Max
Accuracy	<b>Ours</b>	<b>0.724 ± 0.042</b>	<b>0.265 ± 0.044</b>	<b>0.918 ± 0.027</b>
	Orthogonal	0.684 ± 0.048	0.200 ± 0.040	0.832 ± 0.042
	Mix-of-Show	0.659 ± 0.036	0.163 ± 0.037	0.855 ± 0.037
	Prompt +	0.582 ± 0.039	0.102 ± 0.030	0.816 ± 0.039
Identity	<b>Ours</b>	<b>0.731 ± 0.091</b>	<b>0.617 ± 0.083</b>	<b>0.849 ± 0.105</b>
	Orthogonal	0.708 ± 0.142	0.593 ± 0.152	0.828 ± 0.148
	Mix-of-Show	0.666 ± 0.143	0.543 ± 0.158	0.804 ± 0.136
	Prompt +	0.631 ± 0.122	0.519 ± 0.116	0.783 ± 0.132
CLIP-I	<b>Ours</b>	<b>0.722 ± 0.061</b>	<b>0.495 ± 0.074</b>	<b>0.860 ± 0.054</b>
	Orthogonal	0.698 ± 0.066	0.475 ± 0.073	0.834 ± 0.059
	Mix-of-Show	0.652 ± 0.075	0.459 ± 0.069	0.803 ± 0.076
	Prompt +	0.650 ± 0.071	0.455 ± 0.072	0.801 ± 0.061
DINO	<b>Ours</b>	<b>0.510 ± 0.052</b>	<b>0.189 ± 0.088</b>	<b>0.771 ± 0.041</b>
	Orthogonal	0.502 ± 0.101	0.181 ± 0.092	0.749 ± 0.043
	Mix-of-Show	0.412 ± 0.042	0.141 ± 0.051	0.681 ± 0.081
	Prompt +	0.445 ± 0.061	0.137 ± 0.066	0.703 ± 0.105

### 2. Our Approach with SOTA

We extend our implementation to incorporate state-of-the-art (SOTA) methods, such as Orthogonal Adaptation. By integrating our approach, we further enhance the disentanglement capabilities of these methods while mitigating

concept interference. As illustrated in Fig. 1, LoRACLR effectively resolves issues such as Messi’s hair blending with Taylor’s, demonstrating its ability to refine concept separation in challenging cases. These findings highlight LoRACLR’s potential to further improve existing SOTA, making it a valuable complement to current techniques.



<GOSLING> and <MESSI> and <TAYLOR>, in an ancient grand library with towering shelves...

Figure 1. **Combining LoRACLR with SOTA.** Our method can be combined with other methods to reduce interference, resolving cases like Messi’s hair blending with Taylor’s, demonstrating its ability to enhance existing SOTA methods.

### 3. Concept Interactions

Our method is capable of handling interactions between multiple concepts, such as ”holding hands” and ”waving,” ensuring coherent composition and spatial relationships. As shown in Fig. 2, our approach successfully generates complex interactions between subjects while maintaining realism and consistency.



<GOSLING> meets with his friends in a garden and waves his hand at <LEBRON> and <TAYLOR> and <MESSI>...

Figure 2. **Interactions between Concepts.** Examples of subjects holding hands and waving, showcasing our method’s ability to generate coherent multi-concept compositions.



*<LEBRON> is wearing clothes in noir defective style, and <TAYLOR> has the crown on the top of head, and they are on the deck of a wooden ship.*



*<GOSLING> in a denim jacket, holds a basket of apples, and <TAYLOR> is wearing striped shirt and linen trousers, and they are in a farmers market.*



*<GOSLING> is dressed as a medieval knight, <MARGOT> wears a green dress with a mask while holding glowing lanterns and they are in the library.*



*<TAYLOR>, in a traditional samurai outfit, stands beside <MESSI>, who has a mechanical arm with gears, on the edge of a cliff.*

Figure 3. **Diverse Prompt Configurations.** Examples demonstrating the flexibility of our approach in multi-concept generation. Concepts are modified with different attributes, such as styles, objects, and accessories, enabling more controlled and varied outputs.

## 4. Concept Placement Diversity

Our method effectively handles diverse and non-linear concept configurations, enabling flexible spatial arrangements. As shown in Fig. 4, our approach successfully generates compositions where objects are stacked, such as a vase placed on top of a chair, or positioned back-to-back, such as a dog and a cat in front of a pyramid. This demonstrates our model’s ability to preserve spatial relationships while maintaining visual coherence.

## 5. Flexibility in Prompt Setup

We use the same setup as Orthogonal Adaptation for prompts (e.g.,  $\langle V1 \rangle$ ,  $\langle V2 \rangle$ ,  $\langle V3 \rangle$ ...), but our method is flexible and accommodates different prompt configurations. As shown in Fig. 3, our approach allows generating concepts with various objects or accessories, such as “ $\langle V1 \rangle$  in noir style” or “ $\langle V2 \rangle$  with a crown”. This flexibility enables more diverse and controllable multi-concept compositions.

## 6. More Comparison

In addition to the comparisons presented in the main paper, this section highlights further evaluations to emphasize the robustness of our method.

### 6.1. Comparison with OMG

OMG [2] relies on a two-step process for scene generation. First, it generates a layout that structures the composition of the scene. Next, it populates this layout by placing the subjects in their respective positions. This dependency on intermediate layout generation introduces notable limitations. Errors in the layout creation stage often propagate, resulting in inconsistencies in the final output. Additionally, OMG struggles with scenarios involving subjects that share similar attributes, such as two individuals of the same gender (e.g., two women). This limitation leads to reduced quality and coherence in the generated images. Furthermore, since OMG operates in two stages, it requires approximately twice the inference time compared to single-stage approaches, e.g., ours, Mix-of-show and Orthogonal Adaptation, making it less efficient for real-time or large-scale applications.

In contrast, our method bypasses the need for intermediate layouts, directly producing coherent and visually appealing compositions. As shown in Fig. 6, our approach excels in creating realistic and contextually aligned scenes, such as “...on the street, drinking a coffee” and “...in a cool restaurant, delicious meals on the table.” These examples highlight the superior fidelity and contextual understanding achieved by our method compared to OMG [2].



Two  and  and  , in a living room...






 and  , in front of  , during sunset...

Figure 4. **Diverse Concept Placement.** Examples of stacked and back-to-back object configurations, demonstrating our method’s ability to generate flexible spatial arrangements.

## 6.2. More Qualitative Comparison

This subsection provides additional qualitative results to highlight the strengths of our approach in generating multi-concept scenes, from 2 concepts to 6 concepts. Compared to existing methods such as Orthogonal Adaptation [4], Mix-of-Show [1], and  $\mathcal{P}^+$  [5], our method excels in producing coherent, contextually accurate, and visually appealing compositions, even in complex scenarios involving multiple concepts and intricate stylistic requirements.

Figure 7 showcases examples such as “...working in a bustling kitchen preparing a dish with steam rising from pots and pans.” Our method accurately captures the dynamic nature of the scene, ensuring proper interactions between concepts and retaining their distinct identities. In “...inside a futuristic spaceship, sci-fi realism,” the futuristic aesthetics and intricate details are vividly rendered, demonstrating the superiority of our approach in handling complex compositions compared to baselines, which often introduce artifacts or fail to maintain consistency.

Figure 8 further highlights the versatility of our method with scenes such as “...performing a surgery together in an operating room.” Our model not only preserves the realism of the surgical environment but also ensures that all concepts are seamlessly integrated into the scene. In another example, “...investigating a crime scene in noir detec-

tive style,” our method faithfully reproduces the intended stylistic elements while maintaining accurate subject interactions—a challenge for baseline methods that struggle to balance style and coherence.

Finally, Fig. 9 presents challenging scenarios like “...in an ancient grand library with towering shelves.” Our method captures the details of the setting while ensuring the concepts interact naturally within the environment. In “...inside a futuristic spaceship, sci-fi realism,” the vivid rendering of the scene’s futuristic details once again underscores the robustness of LoRACLIR compared to baselines that exhibit inconsistencies in subject placement and stylistic alignment.

## 7. User Study Details

We conducted a user study to evaluate identity preservation and composition quality in generated images. Participants were shown reference images alongside generated scenes (Fig. 5) and asked to rate identity similarity on a scale of 1 (does not look similar) to 5 (looks very similar).

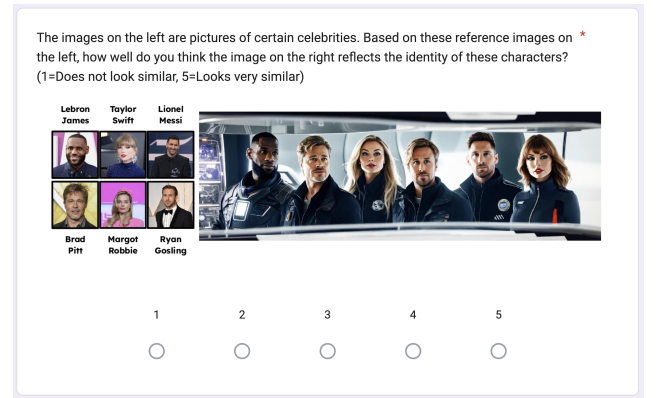


Figure 5. **User Study Interface.** Participants rated identity similarity between reference images and generated scenes, focusing on accuracy and realism.

Ours



OMG



<CHRIS> & <TAYLOR>, on the street, drinking a coffee...

Ours



OMG



<LAWRENCE> & <TAYLOR>, in a cool restaurant, delicious meals on the table...

Figure 6. Comparison between our method and OMG for generating multi-concept scenes. OMG struggles with intermediate layout dependence and compositional errors, particularly with same-gender concepts, while our method achieves seamless and accurate results.



*<LEBRON> & <PITT> & <GOSLING> & <MESSI>, on the deck of a wooden ship, in adventurous fantasy style...*



*<LEBRON> & <MARGOT> & <GOSLING> & <MESSI> & <PITT> & <TAYLOR>, inside a futuristic spaceship, sci-fi realism...*



*<MARGOT> & <GOSLING> & <MESSI>, working in a bustling kitchen preparing a dish with steam rising from pots and pans...*

Figure 7. **Qualitative comparison of multi-concept scenes.** Our method effectively captures dynamic interactions and complex stylistic elements, as seen in examples such as bustling kitchens and futuristic spaceships. It surpasses Orthogonal Adaptation, Mix-of-Show and  $\mathcal{P}+$  in coherence and realism.



Figure 8. **Additional multi-concept image generation examples.** Our method demonstrates superior integration of concepts and themes in diverse scenarios, such as operating rooms and detective noir settings, while maintaining stylistic fidelity.

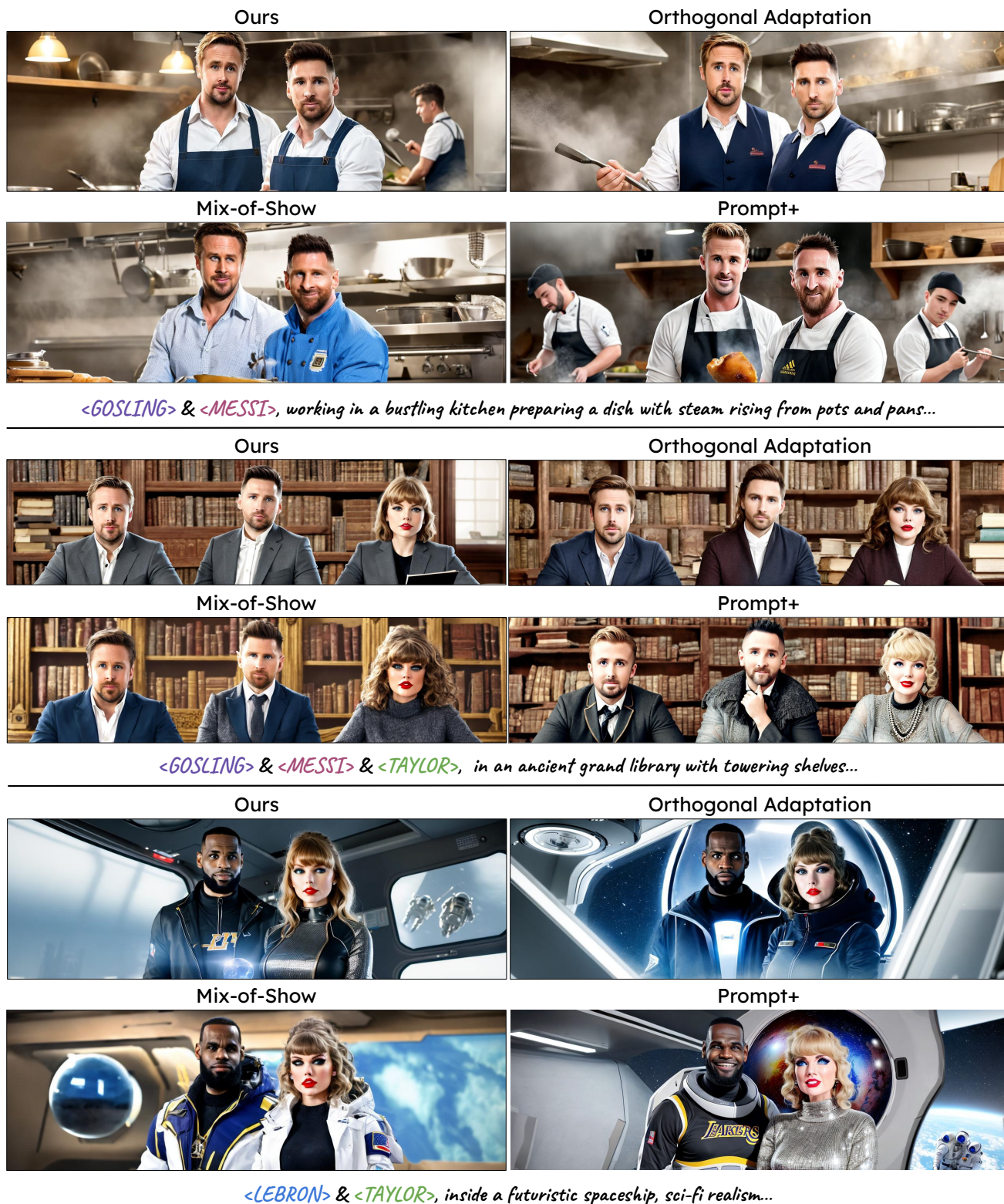


Figure 9. **Extended qualitative results for multi-concept image generation.** It showcases our method’s ability to generate intricate compositions, such as ancient libraries and sci-fi interiors. These results emphasize the robustness of our approach in maintaining style, subject integrity, and contextual relevance.

## References

- [1] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *arXiv preprint arXiv:2305.18292*, 2023. [3](#)
- [2] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhao Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *arXiv preprint arXiv:2403.10983*, 2024. [2](#)
- [3] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Sizhe Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 27948–27959, 2024. [1](#)
- [4] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7964–7973, 2024. [3](#)
- [5] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. [3](#)