Scaling Multi-Identity Consistency for Image Customization via Multi-to-Multi Matching Paradigm

Supplementary Material

6. SIR Scores of OmniGen2

We report the single identity reward (SIR) scores of OmniGen2 [36] with different generation seeds along denoising steps in Figure 8. Similar to the trend of the SIR scores of UNO [38] in Figure 4, the SIR scores of OmniGen2 varies drastically during former steps, *i.e.*, first 10 steps, while getting relatively stable during latter steps. And the generation result with the highest score preserves identity better than that with the lowest score. Also, the UNO's result with the highest score (around 0.8) in Figure 4 preserves identity much better than the OmniGen2's result with the highest score (around 0.4).

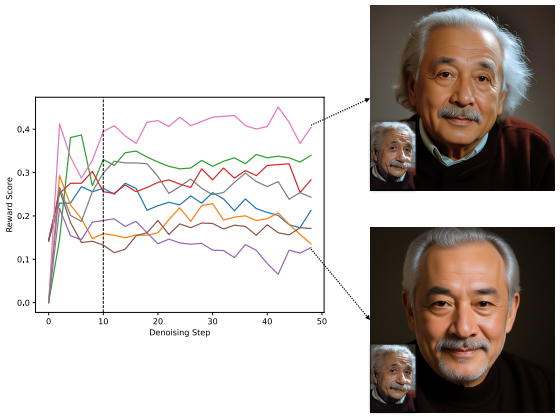

Figure 8. Single identity reward (SIR) scores of OmniGen2 [36] with different generation seeds along denoising steps. The scores become stable after step 10. And the results with highest and lowest reward scores indicating its discriminatory ability.

7. Detailed Quantitative Comparisons

We report detailed comparison on each task type from OmniContext [36] as shown in Table 5, Table 6 and Table 7. In all SINGLE, MULTI and SCENE task types from OmniContext, UMO significantly boosts the ID-Sim and ID-Conf on both pretrained models, *i.e.*, UNO [38] and OmniGen2 [36], leading over previous methods, *e.g.*, MS-Diffusion [34] and DreamO [22]. The comprehensive evaluation demonstrate the effectiveness and generalization of UMO training framework to improve identity consistency and mitigate confusion.

Method	Overall	ID-Sim†	ID-Conf†	AVG
MS-Diffusion [34]	5.83	2.89	6.05	4.92
DreamO [22]	7.65	5.09	5.83	6.19
UNO [38]	6.72	2.11	4.48	4.44
UMO (Ours)	6.77	5.19	7.03	6.33
OmniGen2 [36]	**7.82**	4.75	**7.08**	6.55
UMO (Ours)	7.78	**7.95**	6.72	**7.48**

Table 5. Quantitative results on task type SINGLE from OmniContext.

Method	Overall	ID-Sim†	ID-Conf†	AVG
MS-Diffusion [34]	4.75	2.18	6.97	4.63
DreamO [22]	7.05	4.21	7.12	6.13
UNO [38]	4.48	1.75	5.23	3.82
UMO (Ours)	5.35	4.46	7.20	5.67
OmniGen2 [36]	**7.23**	2.86	6.67	5.59
UMO (Ours)	7.14	**6.61**	**9.04**	**7.60**

Table 6. Quantitative results on task type MULTI from OmniContext.

Method	Overall	ID-Sim†	ID-Conf†	AVG
MS-Diffusion [34]	3.95	1.90	6.75	4.20
DreamO [22]	4.52	4.03	5.74	4.76
UNO [38]	3.59	1.87	5.03	3.50
UMO (Ours)	4.38	4.22	5.58	4.73
OmniGen2 [36]	6.71	2.91	5.31	4.98
UMO (Ours)	**6.78**	**6.65**	**7.03**	**6.82**

Table 7. Quantitative results on task type SCENE from OmniContext.

8. More Qualitative Results

We show more qualitative results on XVerseBench [4] in Figure 14 and Figure 15, and OmniContext [36] in Figure 16 and Figure 17. UMO improves identity similarity without confusion on both single identity and multi-identity scenarios, showing its general and scalable effectiveness.

- **Single Identity**: In Figure 14, UNO [38] itself generates customization results with low identity fidelity. As a comparison, UMO gets much more similar generated identities. In Figure 16, although OmniGen2 [36] gets moderate fidelity, UMO based on it still achieves remark-

able improvement without degradation of subject similarity (*e.g.*, clothes, *etc*) or prompt following. The observation in single identity scenario demonstrates the extraordinary potential of UMO to enhance identity consistency across several existing models.

- **Multi-Identity**: In Figure 16, UNO [38] suffers low similarity of facial features and identity confusion, *e.g.*, the two generated identities in the last row have almost the same facial features which is the "average" facial features of the two reference ones. By contrast, UMO shows its superiority with higher fidelity and without identity confusion. In Figure 17, the results of OmniGen2 [36] show moderate identity similarity, while UMO still boosts it without degradation. The observation in the scenario of multi-identity shows the impressive promoting ability of UMO to improve identity fidelity and alleviate confusion on existing image customization methods.

Figure 9. More diverse examples of UMO with expression changes.

Also, we include more diverse examples in Fig. 9 to demonstrate that UMO can accurately generate images aligned with users' demands for expression changes.

9. More Results of User Study

We report the radar chart of our user study conducted in Section 4.4, as shown in Figure 10. UMO gets significant improvement across all evaluated dimensions, *i.e.*, identity consistency, prompt following, aesthetic, and achieves the best preference among experts and non-experts, demonstrating the effectiveness of *multi-to-multi matching* paradigm. In all the evaluated dimensions, the boosting effect on identity consistency is the most remarkable, indicating that UMO does optimize towards better identity preservation.

10. More Results of Ablation Study

We conduct further ablation study with UMO trained on OmniGen2 [36] on OmniContext [36] in Table 8.

As shown in the first two rows in Table 8, finetuning OmniGen2 with the same data as UMO (*i.e.*, raw SFT) leads to minor improvement especially in ID-Sim and ID-Conf

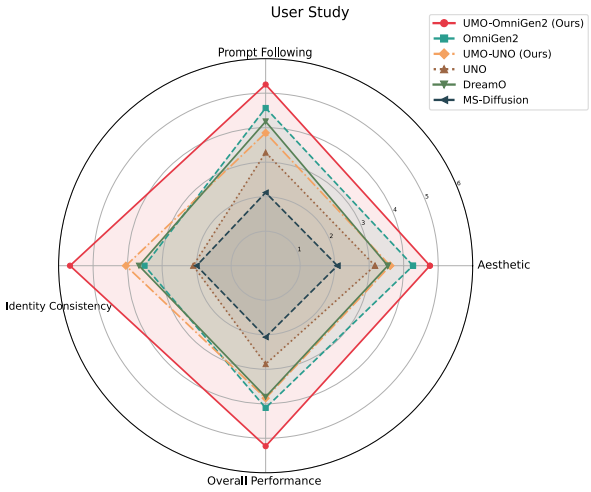

Figure 10. Radar charts of user evaluation of methods on different dimensions.

Method	Overall	ID-Sim†	ID-Conf†	AVG
OmniGen2 [36]	<u>7.23</u>	2.86	6.67	5.59
SFT	**7.24**	3.38	6.80	5.81
ReReFL w/ SIR	7.14	<u>6.44</u>	<u>7.32</u>	<u>6.97</u>
UMO (Ours)	7.14	**6.61**	**9.04**	**7.60**

Table 8. Ablation study with OmniGen2 as the pretrained model on task type MULTI from OmniContext.

scores, while optimizing OmniGen2 with ReReFL demonstrates significant improvement. The comparison indicates the necessity to utilize reinforcement learning with reward focusing on facial region to unleash potential in identity consistency. Instead, vanilla SFT would suppress attention of facial feature due to its small proportion. The comparison is similar to that in Table 4, indicating the generalized effect of ReReFL.

Besides, the last two rows of Table 8 demonstrate that training with SIR instead of MIMR has a significant drop in both ID-Sim and ID-Conf in multi-subject scenario. The comparison proves the effectiveness of MIMR through assigning correct facial supervisions to boost identity consistency and mitigate confusion. Similar observation in Table 4 showing the generalization of the effect of MIMR.

Also, we show the effect of our proposed ReReFL in UMO by comparing simply adding an ID loss during SFT. As shown in Fig. 11, UMO achieves higher ID consistency during all training steps. And the ID-Sim at step 500 of UMO is even higher than that at step 4000 of simply adding ID loss. The proposed ReReFL in UMO is much more effective than simply adding an ID loss.

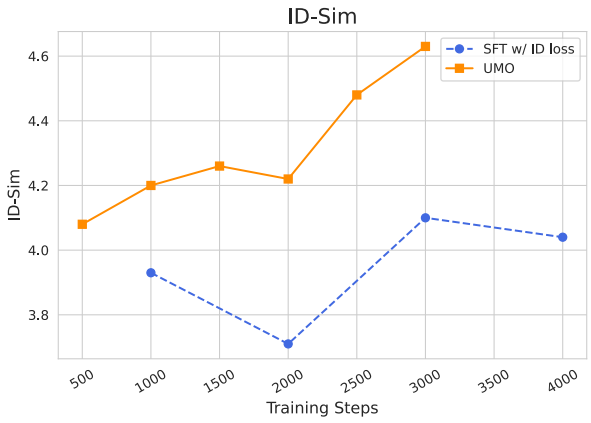

Figure 11. Effectiveness of our proposed ReReFL in UMO.

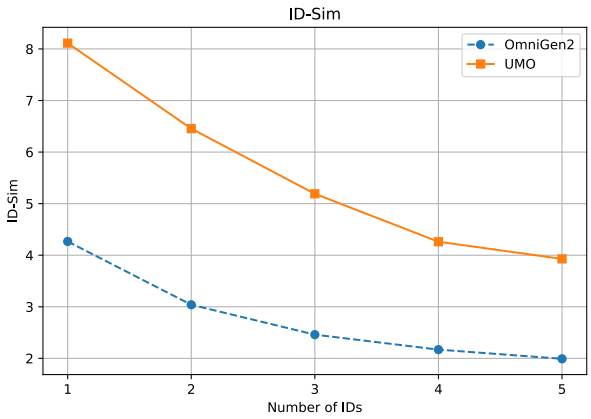

Figure 12. Scalability of UMO.

11. Discussion

- **Robustness**: As shown in Figure 4 and Figure 8, the cosine distance between identity embeddings becomes stable during latter generation steps, indicating that the SIR score has robustness with fewer diffusion steps as perturbations. Since the MIMR score is build upon the optimal assignment with connections weighted by SIR scores between reference identities and generated identities, the robustness of MIMR score is guaranteed by the stability of the SIR scores.
- **Scalability**: Although we build UMO to maintain high-fidelity identity preservation and alleviate identity confusion with scalability to multi-identity, stably scaling to more identities is still restricted due to the dramatic decrease of the pretrained models' reference ability when the number of reference images or identities increases, which demonstrates a similar view with [46].
  As shown in Fig. 12, where we compare UMO with baselines on various IDs setting, as the number of IDs increases, OmniGen2's ID-Sim drops drastically, while UMO achieves much higher performance and maintains a consistent margin over the baseline, demonstrating UMO's stronger ID scalability. However, UMO gets unstable with growing number of IDs, *e.g.*, as Fig. 13. We found that it's blamed to the weak stability of the based model. Given 5 reference IDs, OmniGen2 itself generates an image with only 2 people with low similarity. Though UMO boosts ID consistency of these 2 people with global assignment, it's hard for UMO to add the other 3 reference people. With a more powerful based model, we believe that UMO would achieve greater performance with reference IDs increase.
- **Generality**: UMO proposes multi-to-multi matching paradigm to improve identity consistency and alleviate confusion in multi-ID scenarios, which is more challenging and attracting than non-human objects, because of our sensitivity to human faces. However, the matching paradigm boosts performance not only on human but also on general subjects (IP-Sim) as shown in Tab. 2 in the main paper. We believe UMO would be effective on non-human objects with a comparable fine-grained feature extractor to the one for face recognition.

OmniGen2 UMO

Figure 13. Failure case of UMO.

Figure 14. Qualitative results on task type Single-Subject from XVerseBench [4].

Figure 15. Qualitative results on task type Multi-Subject from XVerseBench [4].

Figure 16. Qualitative results on task type SINGLE Character from OmniContext [36].

Reference Images	**Prompt**	OmniGen2	**UMO** (Ours)
	Please make the man and the women play a computer game together.		
	Please make the woman and the man play chess together.		
	I wish to see the person from figure 1 and the individual from photo2 pointing at the ground together.		
	Have the person run together with the man in a dense forest.		
	Two individuals lie down and rest with eyes closed in a peaceful park.		

Figure 17. Qualitative results on task type MULTI Character from OmniContext [36].