MICo-150K: A Comprehensive Dataset Advancing Multi-Image Composition

Supplementary Material

In the accompanying **zip archive**, we provide the complete set of prompts for MICo-Bench, along with captions for all source images in each case. We also provide the meta-prompts used to query GPT-4o [17] when computing the Weighted-Ref-VIEScore's Perceptual Quality (**PQ**) and Semantic Consistency (**SC**) scores. We release the generation outputs of Qwen-MICo on MICo-Bench, together with their raw evaluation results. Due to the 200 MB submission file size limit, We provide 15 examples for the De&Re task, and three examples per task type for all other types.

In this **Appendix**, we organize the supplementary materials into the following sections:

A1. Reproducibility, Licensing, Data Release

MICo-150K and MICo-Bench (which are strictly non-overlapping—the training set contains no MICo-Bench images) will be fully released upon acceptance. The release will include all source images (or corresponding metadata), all multi-image composition tasks, and the complete set of GPT-based evaluation prompts. All GPT evaluations in the main paper were conducted with `temperature=0`, yielding deterministic outputs and ensuring full reproducibility.

All MICo-150K source images come from publicly released datasets with open licenses, including Subject200k [16], VITON-HD [6], X2I-Subject-Driven [23], Mulan [18], Echo-4o [11], SUN397 [22], and CC12M [4]. The two headshot datasets used for identity filtering, BKM1804/headshot_istockphoto [2] and BKM1804/headshot_pexels_v1 [3], are released under the **Creative Commons Attribution 4.0 (CC BY 4.0) license**. This license permits redistribution, modification, and research use, provided that proper attribution is given.

Our use of these datasets conforms to all licensing and redistribution requirements. No private or restricted data were used, and no personally identifiable images outside openly licensed sources were collected.

A2. Visualizations and Qualitative Examples

This section provides additional visual examples. Fig. A1 presents more cases from the dataset. Fig. A2 illustrates examples from the five open-source models before and after being trained on MICo-150K. Fig. A3 compares examples generated by GPT-Image-1 [12] and Nano-Banana [7].

A3. Analysis of Human-Face Source Images

A3.1. Data Sources, Licensing, and Ethics

Celebrity Faces. All celebrity portraits in MICo-150K originate exclusively from the public X2I-Subject-Driven dataset [23], which provides high-quality subject-centric photographs under a research-permissive license. No additional celebrity images were crawled or collected from the web. To mitigate right-of-publicity and privacy risks, **raw celebrity images will not be redistributed**; only derived metadata and selection indices will be released.

Non-Celebrity Faces. The remaining portraits are drawn from three established datasets commonly used in virtual try-on and portrait-generation research: VITON-HD [6] (research-only license), Headshot iStockPhoto [2] and Headshot Pexels v1 [3] (CC-BY-4.0). The two headshot datasets released under **CC-BY-4.0** explicitly permit redistribution, modification, and research use with proper attribution, as discussed in Sec. A1. All images were used in strict accordance with their respective licenses.

Across all sources, MICo-150K contains **no private, scraped, or user-uploaded photographs**. The portraits are used solely for constructing multi-image composition tasks; no identity recognition, biometric profiling, or sensitive attribute inference is performed.

A3.2. Demographic Distribution Analysis

To understand the demographic characteristics and potential bias of the human-face source images used in MICo-150K, we employed **Qwen2.5-VL-72B** to estimate three attributes for each portrait: (1) ethnicity, (2) gender, and (3) coarse age group. In total, we analyzed **53,648** human-face source images, including **11,677** images used in the De&Re task and **41,971** images used in all other MICo tasks.

Ethnicity. The estimated ethnic distribution is *East/Southeast Asian: 6,740*, *South Asian: 3,337*, *European/Middle Eastern: 33,982*, *African: 9,290* and *Others: 299*. This reveals an over-representation of European/Middle Eastern subjects and relative under-representation of Asian groups, reflecting the typical skew of widely used portrait datasets.

Gender. The raw gender distribution is *Male: 21,241* and *Female: 32,407*. During MICo-150K task construction, however, we explicitly enforced gender balancing. Across the 159,091 human-face images sampled for all composi-

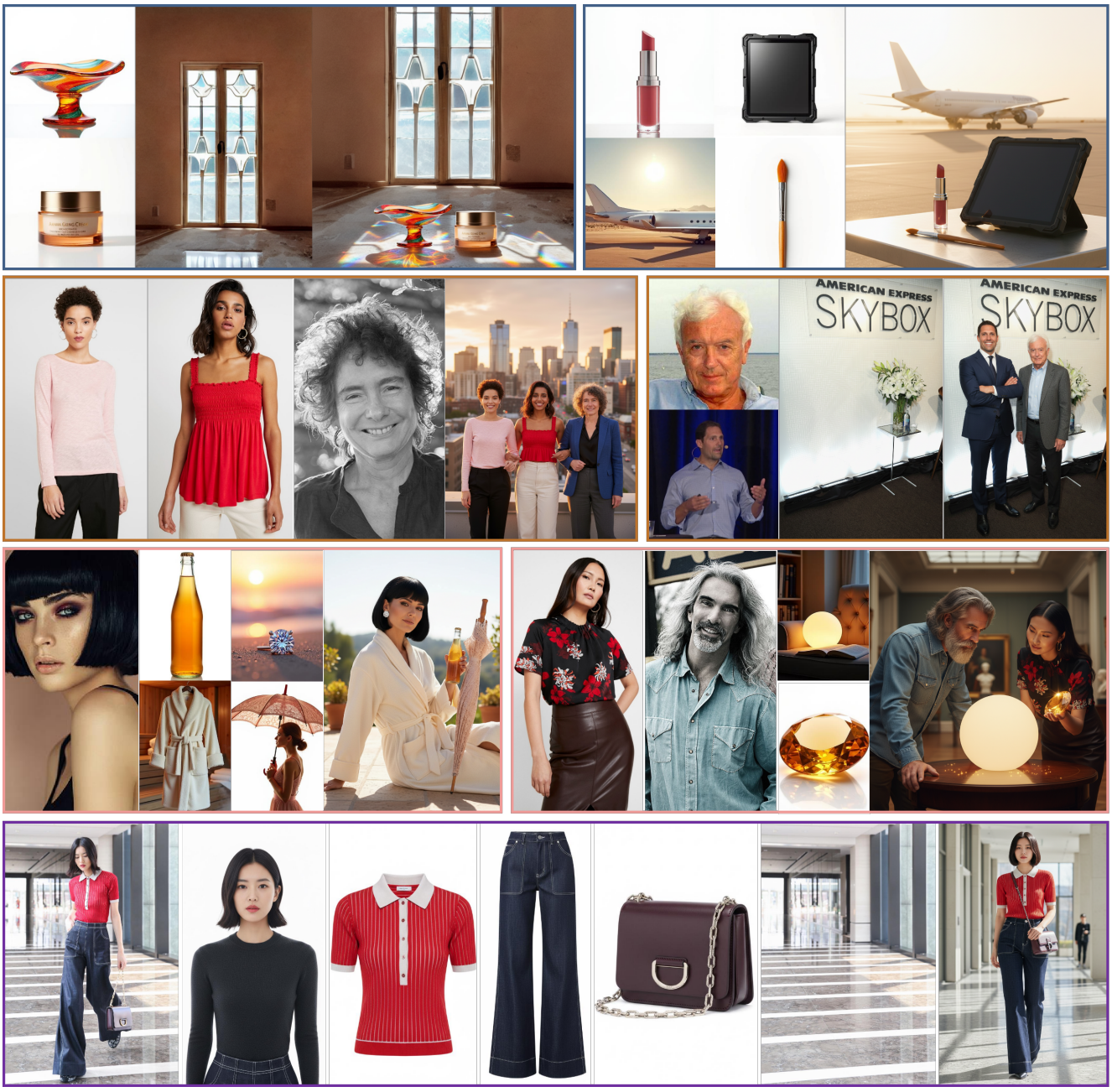


Figure A1. Visualization examples from the MICo-150K dataset. **Row 1** (Object-Centric): "2 objects + scene" and "4 objects" compositions. **Row 2** (Person-Centric): "3 women" and "2 persons + scene". **Row 3** (Human-Object Interaction): "1 person + 4 objects" and "2 persons + 2 objects". **Row 4** (De&Re): the first image is a real-world photo, the last is the recomposed result, with intermediate visual elements including decomposed persons, objects, clothes, and scene components.

tions (see Tab. 1 in the main paper), the resulting ratio becomes approximately 1:1, mitigating upstream imbalance.

Age. The age distribution is *Child: 1,421*, *Teen/Young Adult: 31,413*, *Middle-Aged: 19,146*, and *Older Adult: 1,668*, showing a strong dominance of young adults, a pattern that is common in fashion, try-on, and aesthetic portrait datasets. Importantly, for multi-image composition tasks, this age skew is unlikely to negatively impact model behavior, as real-world applications naturally exhibit a similar young-adult majority. Thus, while the distribution is uneven, it is largely aligned with practical usage scenarios.

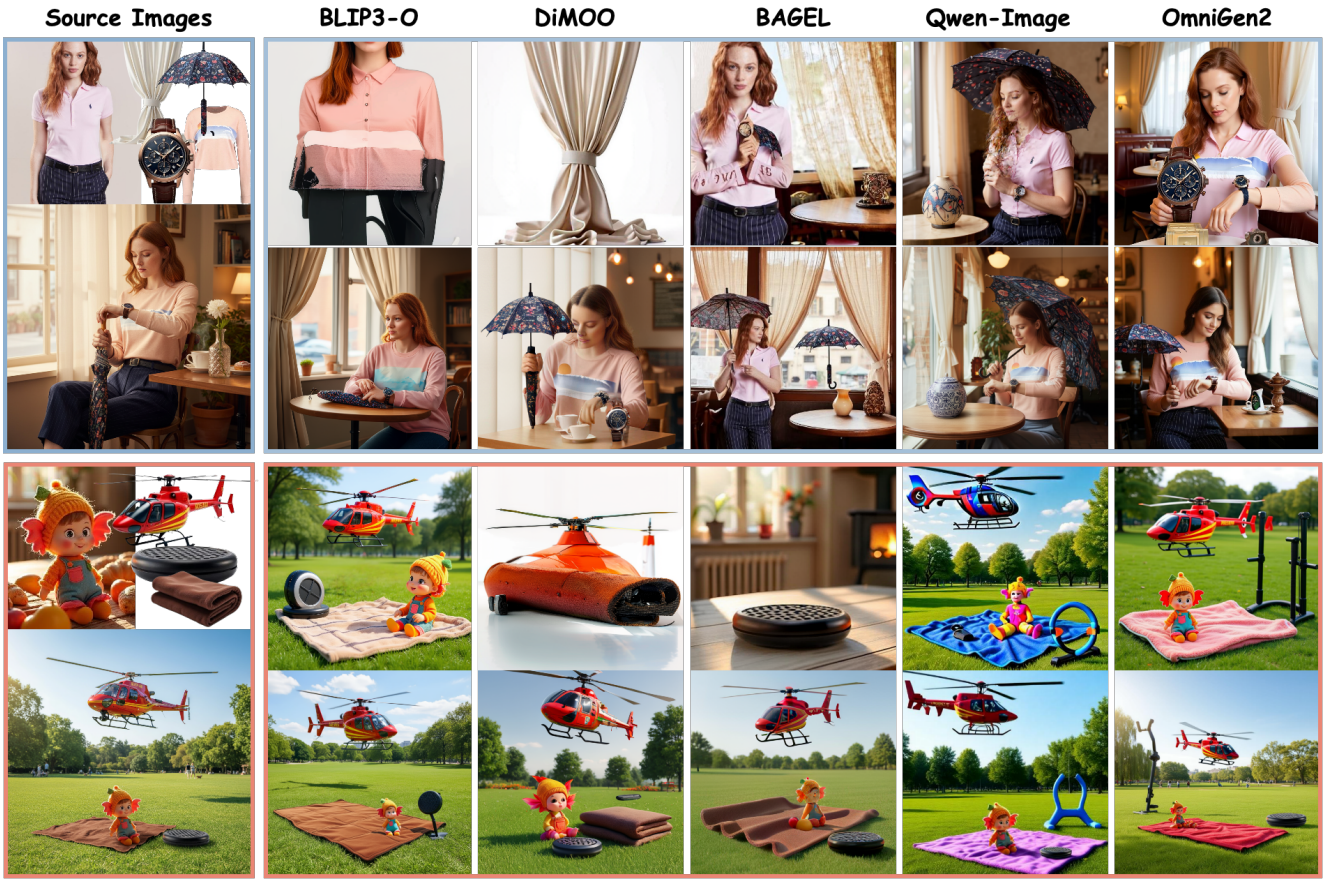


Figure A2. Comparison of open-source models before and after MICo-150K training. Some source images were cropped or background-removed for visualization. BLIP3-o [5] and Luminia-DiMOO [24] gain strong multi-image composition abilities after training. Qwen-Image-Edit [20] and BAGEL [8] were not explicitly trained for MICo tasks, but exhibit emergent MICo capabilities that are further enhanced through fine-tuning. OmniGen2 [21] preserves identity well and produces more aesthetic, prompt-aligned results after training.

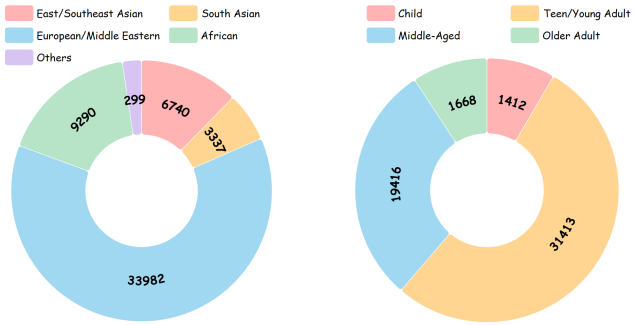


Figure A4. Human-face source images exhibit a Western-centric ethnicity skew and a young-adult bias, reflecting characteristics inherent in the upstream public datasets rather than biases introduced by MICo-150K itself.

Discussion and Mitigation. The demographic analysis highlights several imbalances, particularly a Western-centric ethnicity skew and a young-adult bias. These limitations stem from upstream public datasets and are not specific to MICo-150K. We emphasize that MICo-150K is **not designed** for fairness benchmarking or identity-sensitive evaluations. Instead, its purpose is restricted to studying visual multi-image composition.

Overall, while demographic skew is present, as shown in Fig A4, we openly document these statistics to facilitate informed and responsible use. We caution against applying MICo-150K in fairness-critical or identity-sensitive contexts. Future work may incorporate more diverse portrait sources to improve global representativeness.

A4. Details of Weighted-Ref-VIEScore

A4.1. Weights

As described in Sec. 4.2 of the main paper, the Weighted-Ref-VIEScore applies a multiplicative visibility coefficient W, computed as the ratio between (i) the number of source images whose key content is successfully preserved in the generated target image and (ii) the total number of source

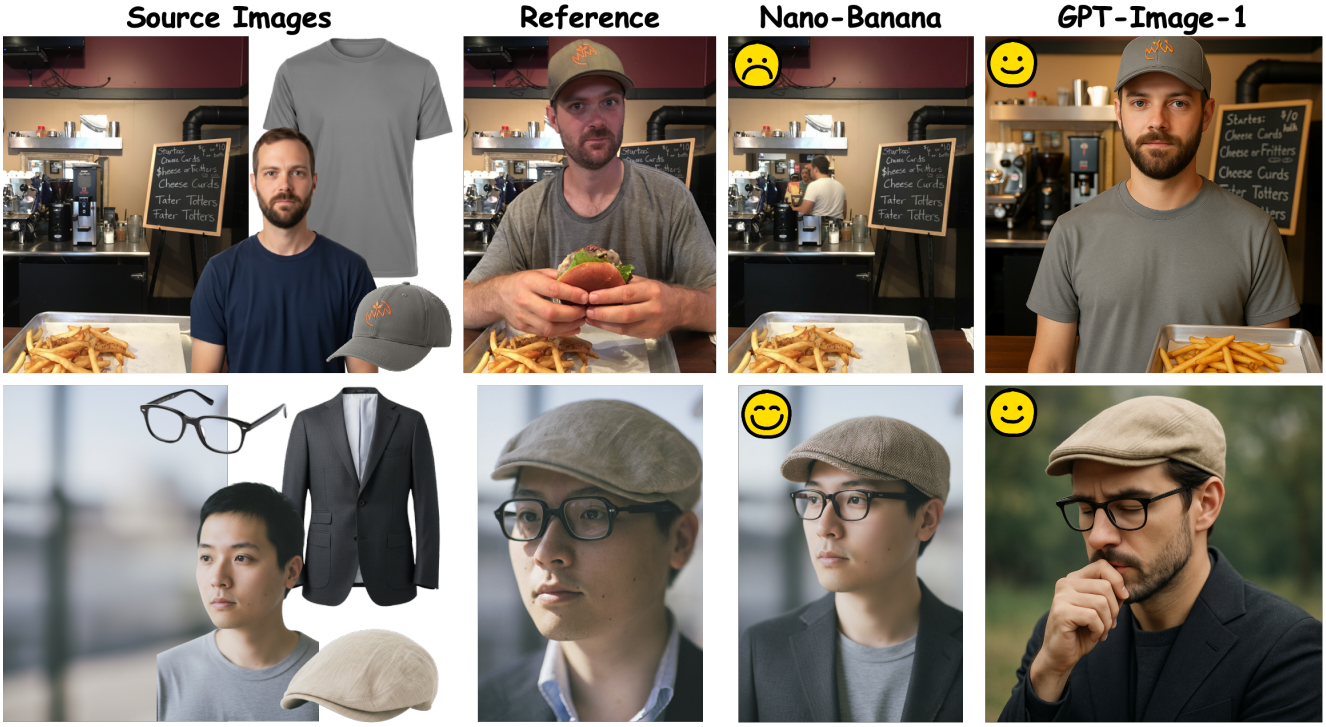


Figure A3. Nano-Banana [7] produces more realistic images with stronger fidelity to the source inputs and a higher quality ceiling, but occasionally fails on certain cases. GPT-Image-1 [12] exhibits a more stylized, less photo-realistic look, yet remains highly stable and consistently yields semantically coherent results.

images. The critical step is therefore determining whether a source element is present in the target image.

Objects, clothing items, and scenes. For non-human content, we employ Qwen2.5-VL-72B [1] to compare the source and target images. For each case, we construct a task-specific binary question (*e.g.*, *"Does the <object> in the source image appear in the target image?"*) and treat the model's answer as a binary indicator of presence.

Human faces. For source images containing faces, we use ArcFace [9] to extract identity embeddings from both source and target portraits and compute cosine similarity. According to the finding of Xu *et al*. [25], face-ID similarity is *not* a "higher-is-better" metric: extremely high similarity often indicates copy–paste artifacts and leads to unnaturally rigid identity transfer. Real photos of the same person typically form a similarity distribution peaking around ~ 0.58, rather than near 1.0. To balance identity consistency with realism, we use task-dependent thresholds: identity is treated as preserved when the similarity exceeds *0.50 for person-centric tasks*; and *0.45 for all other tasks*, where facial details are less essential. These thresholds prevent the metric from favoring trivial copy–paste solutions while still recognizing correct identity transfer in multi-image composition.

A4.2. SC and PQ Scoring

To compute the Semantic Consistency (**SC**) and Perceptual Quality (**PQ**) components of the Weighted-Ref-VIEScore, we query GPT-4o twice, once for SC and once for PQ.

For SC scoring, the model takes two images as input, a reference image and the image to be evaluated, and produces two sub-scores:

- **Prompt Following (PF):** how well the generated image follows the textual input prompt;
- **Subject Resemblance (SR):** how well the generated image matches the reference image in terms of people, scenes, clothing, and objects.

Both PF and SR are scored on a $[0, 10]$ scale, and the final SC score is defined as $\text{SC} = \sqrt{\text{SR} \times \text{PF}}$. We use a fixed meta-prompt for SC evaluation. The full prompt template is provided in the supplementary file `SC_meta_prompt.txt`.

For PQ scoring, the model takes only the image to be evaluated as input and produces two sub-scores:

- **Naturalness:** evaluates whether the generated image appears visually plausible, including coherent lighting, shadows, geometry, and overall realism.
- **Artifacts:** measures the absence of visual defects such as distortions, duplicated limbs, or unnatural textures.

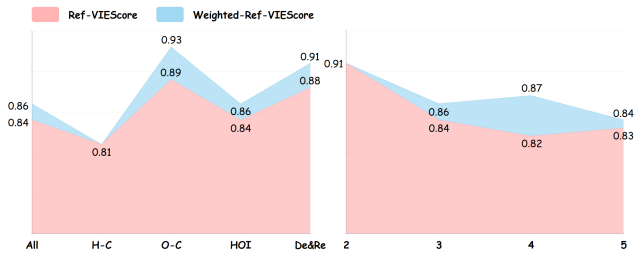


Figure A5. We refer to the metric without the weighting factor W as **Ref-VIEScore**. We conduct extensive human studies and compute the *Spearman rank correlation* between human preferences and the rankings produced by both **Weighted-Ref-VIEScore** and **Ref-VIEScore**. The results are summarized in the figure.

Both naturalness and artifacts are scored on a [0,10] scale, and the final PQ score is defined as their geometric mean. The full prompt template is provided in the supplementary file `PQ_meta_prompt.txt`.

A5. Human Study and Metric Alignment

To validate the effectiveness of Weighted-Ref-VIEScore, we conducted an extensive human study to measure its alignment with human preferences. From the 27 task types in MICo-Bench, we sampled two cases per type, and additionally selected 21 cases from the De&Re task, yielding a total of 75 evaluation cases.

For each case, we randomly selected five candidate outputs from a pool of twelve models: five open-source models (BLIP-3o [5], Lumina-DiMOO [24], BAGEL [8], Qwen-Image-Edit [20], OmniGen2 [21]) in both their pre-training and MICo-finetuned versions, and two closed-source models (GPT-Image-1 [12] and Nano-Banana [7]). Human evaluators were provided with the *source images*, *text prompts*, and the *five anonymized candidate outputs* for ranking.

We recruited 25 human participants, all holding at least a bachelor's degree, including 9 senior Ph.D. students. We further define Ref-VIEScore as the variant of our metric *without* the weighting factor W, *i.e.*, SC $\times$ PQ.

For each case, we computed the **Spearman rank correlation** between each participant's ranking and the rankings produced by Weighted-Ref-VIEScore and Ref-VIEScore. The average correlation across 25 participants was used as the final correlation score for that case. We report results under two groupings: (1) grouped by **task type**, and (2) grouped by the **number of input source images**. The results, summarized in Fig. A5, reveal three key findings:

1. **Strong human–metric alignment.** Weighted-Ref-VIEScore achieves consistently high Spearman correlations across all task types, indicating robust agreement with human judgment.
2. **Robustness to varying numbers of source images.** Thanks to our reference-image design, Weighted-Ref-VIEScore maintains stable accuracy regardless of how many source images are provided as input.
3. **Importance of the weighting factor W.** Weighted-Ref-VIEScore outperforms Ref-VIEScore across all settings. Moreover, the inclusion of W provides interpretability by revealing which source elements failed to appear in the generated image, a capability essential for diagnosing and improving model behavior.

A6. Additional Experiments

A6.1. Reference Image and Style Homology Risk

A natural concern is whether Weighted-Ref-VIEScore may implicitly favor images whose visual style is closer to that of the reference generator, since reference images for most MICo-Bench tasks are synthesized using Nano-Banana [7]. One might hypothesize that models producing Nano-Banana–like aesthetics could receive higher scores.

We first clarify the setup: for all De&Re task cases, the reference images are real photographs, whereas only the reference images for the remaining 27 task types are generated by Nano-Banana.

To evaluate whether such style homology affects scoring, we randomly sampled **108 cases** from the non-De&Re task types (4 cases per type). For each case, we generated outputs from three different models, GPT-Image-1 [12], Qwen-MICo, and Nano-Banana [7]. We then computed a style distance between each model output and the corresponding reference image using DINOv2 [13] ViT-b cosine distance, a strong self-supervised perceptual feature encoder widely used for measuring cross-image similarity.

This produced a sequence of model rankings induced by style distance and a sequence of rankings induced by Weighted-Ref-VIEScore. We computed the Spearman correlation ρ between these two rankings for each case.

Across all 108 cases, the mean correlation was $\rho = 0.251$, indicating only a weak association between reference-style similarity and VIEScore ranking. In other words, models whose outputs resemble the Nano-Banana reference style are **not** systematically favored. Weighted-Ref-VIEScore does **not** reward stylistic imitation, and its scoring behavior is largely orthogonal to reference similarity. Qualitative examples illustrating this lack of dependency are shown in Fig. A6.

A6.2. GPT-4o Evaluator-Generator Coupling

Since our evaluation relies on GPT-4o [17] to compute both the **PQ** (Perceptual Quality) and **SC** (Semantic Consistency) scores, a natural concern is the potential risk of *evaluator–generator coupling*: if a generative model produces images whose style is similar to that of GPT-Image-1 [12] (or when GPT-Image-1 itself is being evaluated), could such outputs receive artificially inflated scores? Although prior

Table A1. We evaluate on the MICo-Bench subset where each case contains exactly three input images, since Qwen-Image-2509 does not support higher-order composition. Qwen-MICo consistently outperforms Qwen-Image-2509 across nearly all evaluation dimensions.

Method	Object Centric		Person Centric					HOI			
	2O1S	3O	2M1W	2W1M	3M	3W	2P1S	1P2O	2P1O	1P2C	1P1C1O
Qwen-MICo	**56.12**	**59.56**	**59.04**	**58.96**	**50.11**	**56.19**	**60.97**	**54.92**	52.16	**55.82**	**54.26**
Qwen-Image-2509	56.00	45.32	42.63	52.46	48.40	50.78	49.70	50.64	**54.65**	51.77	47.91

Figure A6. Similarity to the reference image does not influence how Weighted-Ref-VIEScore ranks model outputs, demonstrating that the metric remains objective and well-aligned with human preferences.

works [10, 19] suggest that GPT-based evaluators do *not* inherently favor GPT-generated content, we explicitly examine this possibility in the MICo evaluation setting.

Following the procedure in Sec. A6.1, we sampled the same 108 non-De&Re cases from MICo-Bench and obtained outputs from three models: GPT-Image-1 [12], Qwen-MICo, and Nano-Banana [7]. For each case, we computed a style distance by measuring the DINOv2 feature distance between each model's output and the output generated by GPT-Image-1 for the same case.

We then computed the correlation between this style-distance sequence and three evaluation-score sequences: SC scores, PQ scores, and Ref-VIEScore (SC $\times$ PQ). The resulting Spearman correlations were: $\rho_{\text{SC}} = -0.115$, $\rho_{\text{PQ}} = -0.008$, $\rho_{\text{Ref-VIEScore}} = -0.077$.

All correlations are near zero, indicating that neither SC, PQ, nor Ref-VIEScore favors outputs stylistically closer to GPT-generated images, even though GPT-4o is used as the evaluator. This confirms the absence of evaluator-generator coupling in the MICo evaluation setting. Representative qualitative examples are shown in Fig. A7.

A6.3. Resistance to Copy-Paste Hack

A common failure mode for composition models is the *copy-paste hack*: directly cutting objects or faces from the source images and pasting them onto a new background. An effective metric must not reward such behavior.

To test whether Weighted-Ref-VIEScore can be fooled by this shortcut, we constructed explicit copy-paste baselines. For two representative sub-tasks, **Object+Scene** and

Figure A7. Although PQ and SC scores are computed using GPT-4o, we find no evidence of evaluator–generator coupling: images that are stylistically closer to GPT-Image-1 do *not* receive higher scores. The evaluation depends solely on MICo output quality and remains well aligned with human preferences.

Person+Scene, we randomly sampled 10 cases each. For every case, we segmented the source objects (or person) using an off-the-shelf segmentation tool [15] and manually pasted them onto the scene image, creating a naive composite without any harmonization.

Although these copy-paste outputs achieve a perfect weight factor $W = 1.00$ (since all source elements appear in the target), their **PQ** and **SC** scores are extremely low. Across all 20 constructed cases, the final Weighted-Ref-VIEScore averages only **14.16**, demonstrating that the metric effectively penalizes unnatural compositing and cannot be exploited by trivial cut-and-paste strategies. Representative examples are shown in Fig. A8.

A7. Quantitative Evaluation of Qwen-MICo

For fair comparison, we clarify that Qwen-Image-2509 [14] was released before MICo-150K became available, and thus it could not have used our dataset for training. This eliminates the possibility of data leakage or overlap between MICo-150K and Qwen-Image-2509's training corpus.

To further quantify the performance gap between Qwen-Image-2509 [14] and Qwen-MICo, we evaluate both models on the subset of MICo-Bench containing tasks with exactly three input images (see Tab. 1 in the main paper). This subset includes 11 task types, *2O1S, 3O, 2M1W, 2W1M, 3M, 3W, 2P1S, 1P2O, 2P1O, 1P2C, 1P1C1O*. For each type, we randomly sample five cases, yielding a total of 55 evaluation instances. We then compute the proposed Weighted-Ref-VIEScore for both models across all cases.

As reported in Tab A1, Qwen-MICo consistently outperforms Qwen-Image-2509 across nearly all evaluation dimensions, despite using two to three orders of magnitude less training data and supporting arbitrary numbers of input images, whereas Qwen-Image-2509 is limited to three-image composition. Additional qualitative comparisons are shown in Fig. A9, illustrating the superior compositional fidelity and visual coherence achieved by Qwen-MICo.

In addition, we observe that Qwen-MICo exhibits several remarkable emergent capabilities, with representative examples shown in Fig. A10–Fig. A14.

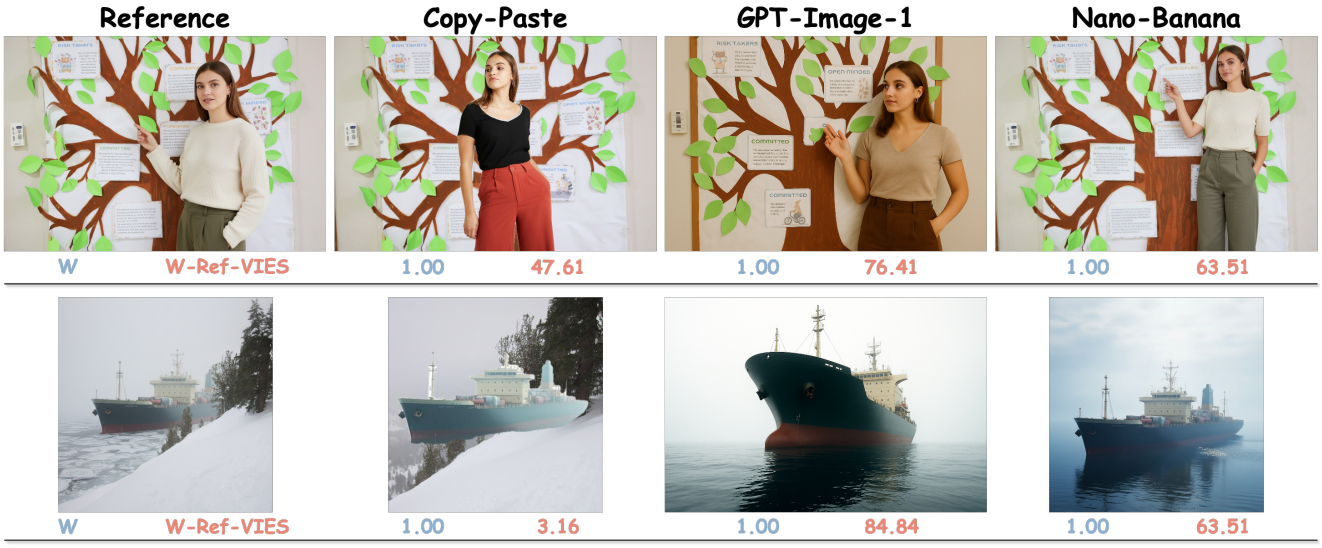
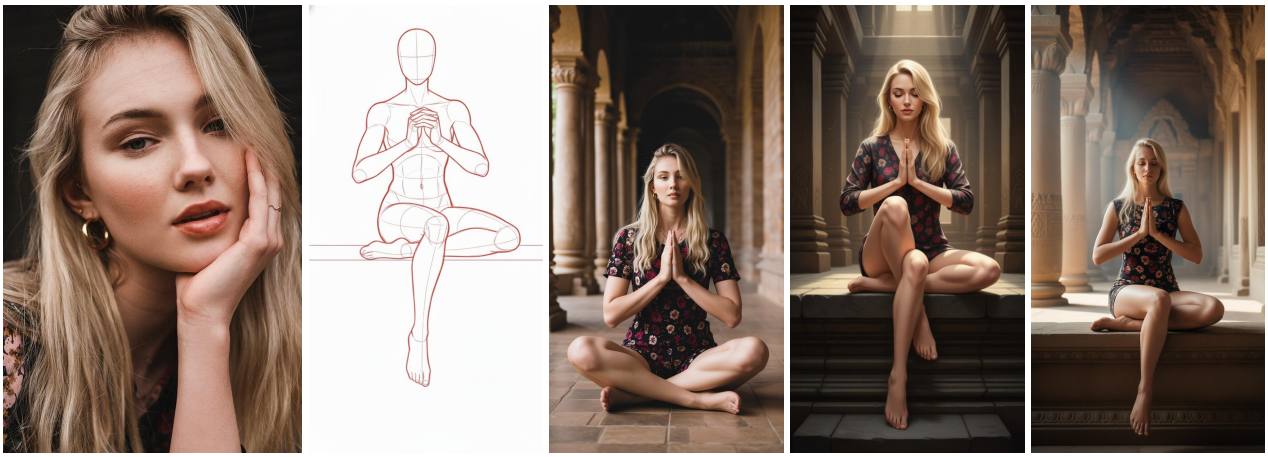


Figure A8. Weighted-Ref-VIEScore effectively **prevents copy–paste hacks**. We segmented objects or persons from the source images and manually pasted them onto the scene image to form a naïve, unharmonized composite. Although such copy–paste results achieve a perfect weight factor (since every source element appears in the output), their PQ and SC scores remain very low.

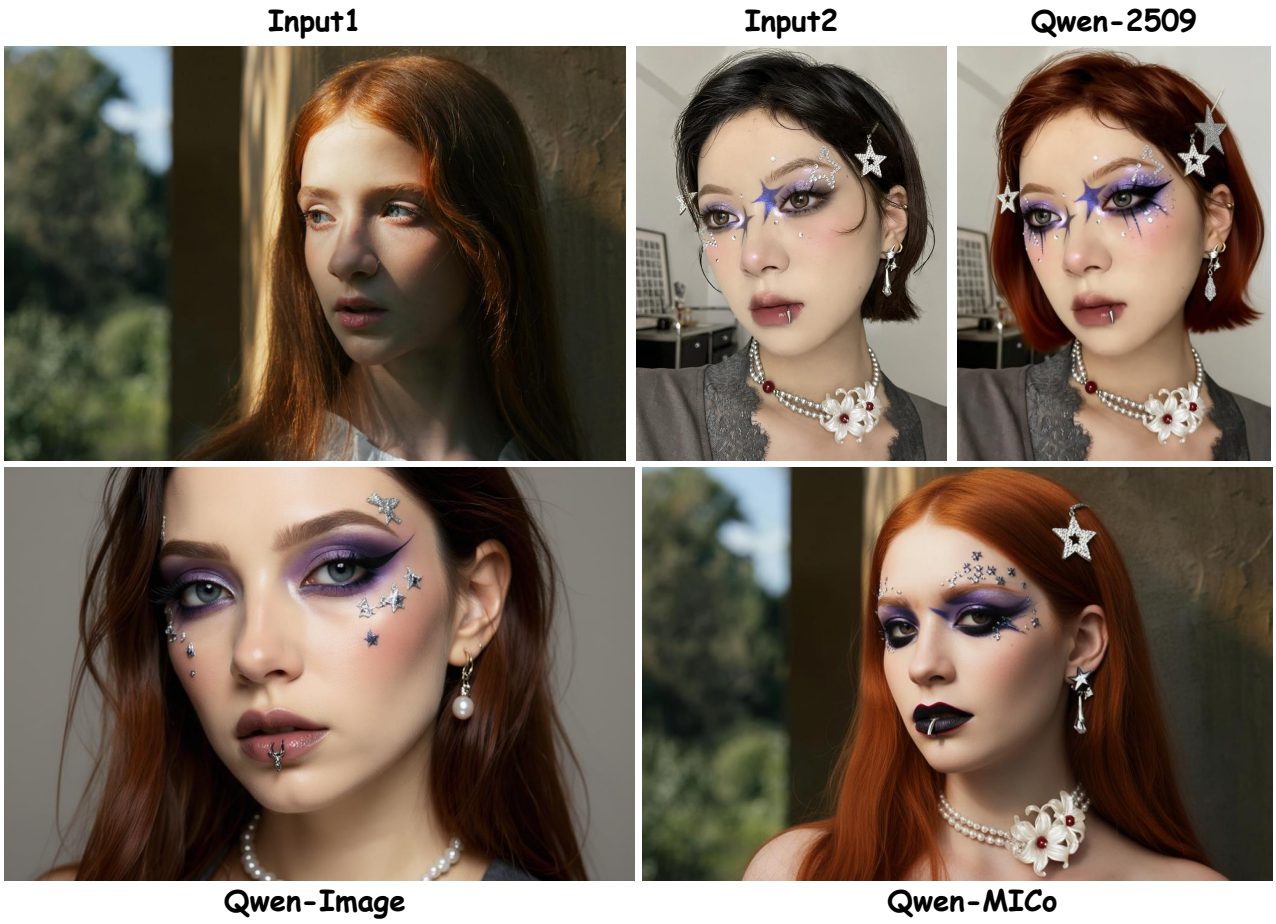
Figure A9. Qwen-MICo consistently outperforms Qwen-Image-2509 across nearly all evaluation dimensions on the MICo-Bench three-image subset. While Qwen-Image-2509 is trained on a massive corpus but restricted to three-image inputs, Qwen-MICo—trained only on MICo-150K—supports arbitrary multi-image composition and yields higher compositional fidelity and visual quality.

Prompt: The man from image 1 is doing the pose from image 2. He is leaning casually on a railing or balcony. His weight rests on his straight left leg, with his right knee bent and the foot resting on the inside of his standing leg. His left arm is relaxed, gripping the rail for support. The right hand is raised to his face, touching his chin in a thoughtful or contemplative gesture. This pose conveys a sense of relaxed elegance and pause. The scene is set against a vibrant city street during the warm hues of dusk, suggesting a moment of quiet reflection.



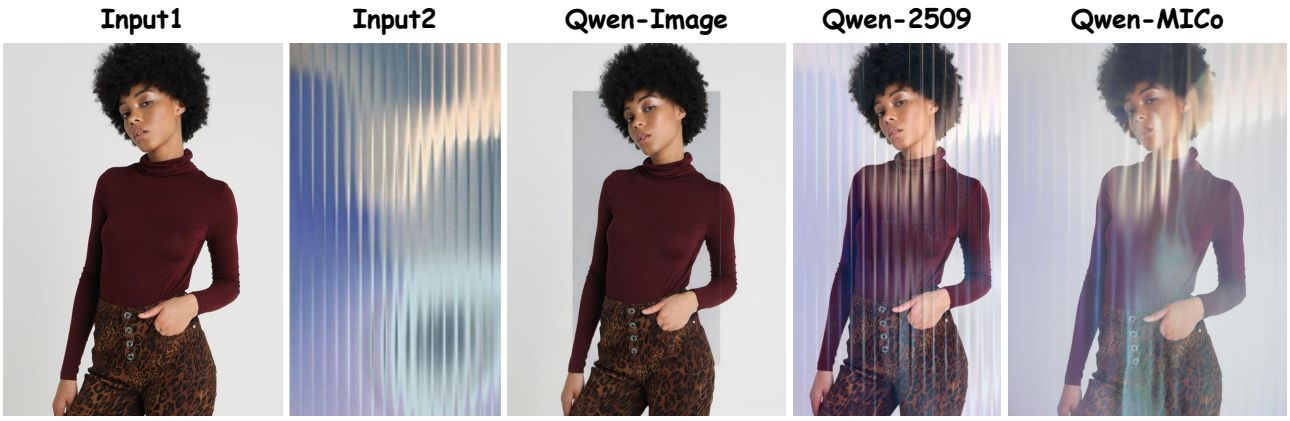
Prompt: The girl from image 1 is doing the pose from image 2. She is seated in a calm, centered pose, suggesting deep contemplation or meditation. She is sitting on an elevated surface with her left leg crossed and tucked towards her body, while her right leg hangs down vertically. Her hands are clasped together at chest level in a prayer-like or focused gesture. The pose is very stable and symmetrical, conveying serenity and quiet focus. This tranquil scene unfolds within the peaceful confines of an ancient temple, where soft, diffused light filters through the grand architecture, enhancing the sense of spiritual quietude.

Figure A10. Qwen-MICo exhibits strong emergent abilities in recognizing and composing complex **human poses**.



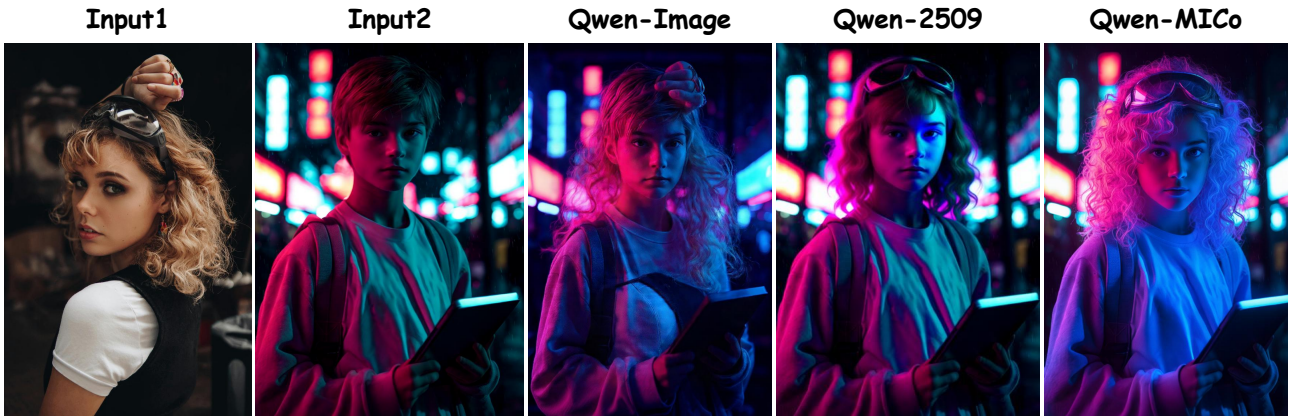
Prompt: The girl from image 1, with her striking red hair, is now adorned with the elaborate makeup from image 2. Her eyes feature lightly applied purple and black eyeshadow, accented by glitter and star-shaped appliqués, creating a bold, artistic look. A lip piercing, a pendant earring, star-sticker on the hair, and pearl-necklace with a floral motif complete her transformation, presenting her in a striking, avant-garde style that contrasts beautifully with her natural features.

Figure A11. Qwen-MICo performs well on **virtual makeup try-on** (transferring the makeup in Image 2 onto the girl in Image 1).



Prompt: Overlay the shimmering, ethereal glass effect from Image 2 onto the photo in Image 1. This transformation will introduce a layer of transparent distortion and refracted light, giving the original image a dreamlike, almost liquid quality. The effect should subtly bend and fragment the underlying visuals, creating an artistic and surreal interpretation that captivates the viewer.

Figure A12. Qwen-MICo shows excellent performance on visually complex tasks that demand **a deep understanding of lighting and optics**, and produces outputs with strong aesthetic appeal.



Prompt: The woman from image 1, with her curious gaze and goggles pushed up, is now bathed in the intense, neon-drenched cyberpunk lighting from image 2. Her curly blonde hair and expressive face are dramatically highlighted by vibrant cyan and magenta hues, casting a futuristic glow upon her while her original dark, workshop-like background remains. This striking contrast creates a captivating visual, blending her grounded essence with an electric, dystopian ambiance.

Figure A13. Qwen-MICo preserves the subject's identity while accurately modeling **lighting and shading**, transferring the illumination from Image 2 onto the girl in Image 1.

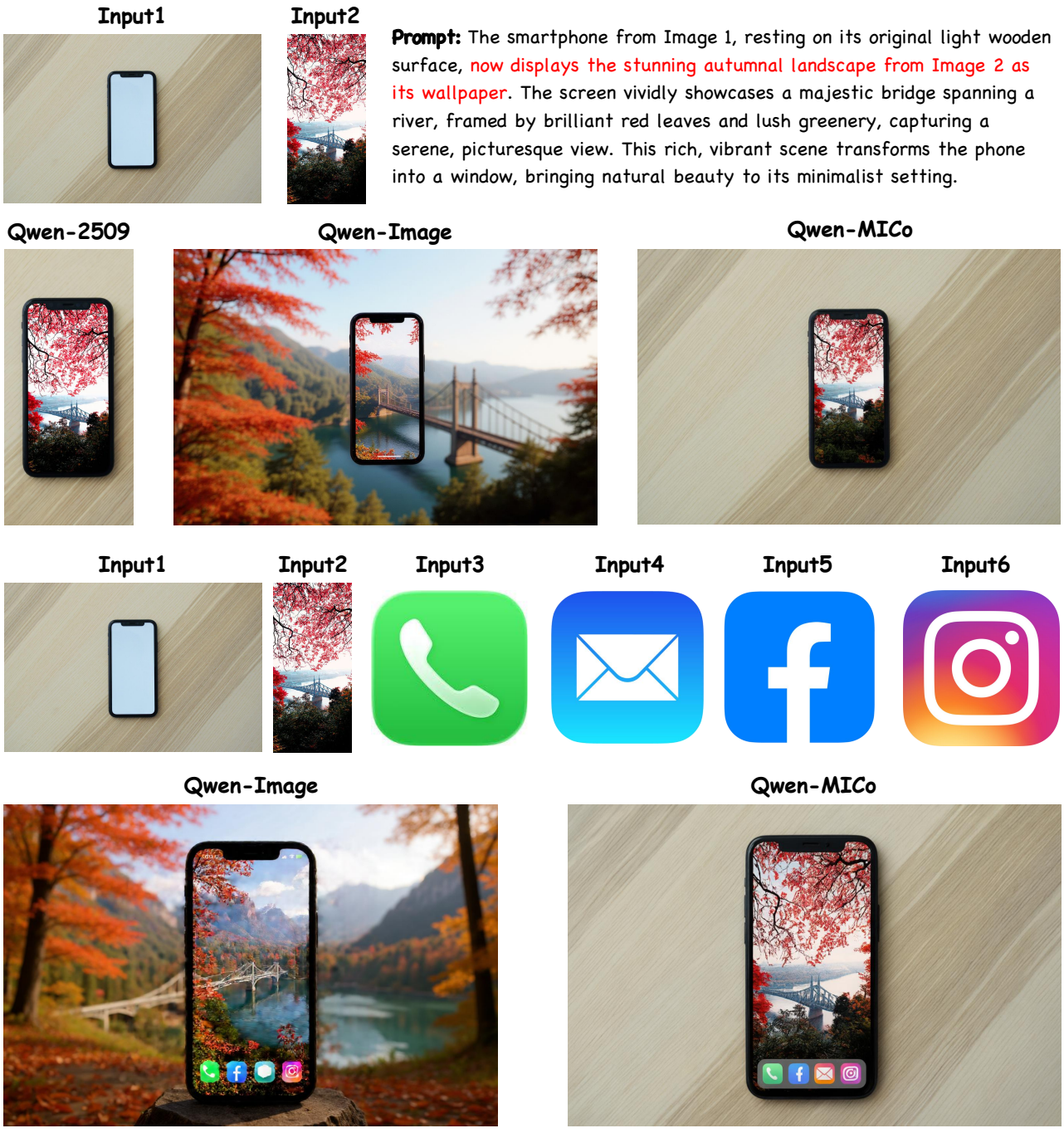


Figure A14. Qwen-MICo preserves the entire appearance of Input 2 while correctly interpreting the prompt phrase *"resting on its original light wooden surface"*. On top of this, it supports more image inputs than Qwen-Image-2509, and accurately renders all application icons onto the phone's home screen.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4

[2] BKM1804. headshot_istockphoto: Headshot image dataset from istockphoto. https://huggingface.co/datasets/BKM1804/headshot_istockphoto, 2025. Accessed: 2025-11-05, 6 000 images. 1

[3] BKM1804. headshot_pexels_v1: High-resolution headshot dataset from pexels. https://huggingface.co/datasets/BKM1804/headshot_pexels_v1, 2025. Accessed: 2025-11-05, includes 3 000 images. 1

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. 1

[5] Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 3, 5

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021. 1

[7] Google DeepMind. Introducing gemini 1.5 flash: Fast, efficient, and multimodal. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2024. Accessed: 2025-10-18. 1, 4, 5, 6

[8] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 3, 5

[9] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. 4

[10] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024. 6

[11] Zihao Wang Leqi Zhu Zhenghao Hu Zilong Huang Jun He Zhiyuan Yan Jinghua Yu Hongsheng Li Conghui He Weijia Li Junyan Ye, Dongzhi Jiang. Echo-4o: Harnessing the power of gpt-4o synthetic images for improved image generation. *https://arxiv.org/abs/2508.09987*, 2025. 1

[12] OpenAI. Gpt-4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. Accessed: 2025-10-18. 1, 4, 5, 6

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5

[14] QwenLM. Qwen-image-edit-2509. https://huggingface.co/Qwen/Qwen-Image-Edit-2509, 2025. Accessed: 2025-11-12. 7

[15] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 7

[16] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. 1

[17] OpenAI Team. Gpt-4o system card, 2024. 1, 5

[18] Petru-Daniel Tudosiu, Yongxin Yang, Shifeng Zhang, Fei Chen, Steven McDonagh, Gerasimos Lampouras, Ignacio Iacobacci, and Sarah Parisot. Mulan: A multi layer annotated dataset for controllable text-to-image generation, 2024. 1

[19] Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025. 6

[20] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 3, 5

[21] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Exploration to advanced multimodal generation, 2025. 3, 5

[22] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010. 1

[23] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation, 2024. 1

[24] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, Jin-

bin Bai, Qian Yu, Dengyang Jiang, Yuandong Pu, Haoxing Chen, Le Zhuo, Junjun He, Gen Luo, Tianbin Li, Ming Hu, Jin Ye, Shenglong Ye, Bo Zhang, Chang Xu, Wenhai Wang, Hongsheng Li, Guangtao Zhai, Tianfan Xue, Bin Fu, Xiaohong Liu, Yu Qiao, and Yihao Liu. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding, 2025. 3, 5

[25] Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Withanyone: Towards controllable and id consistent image generation, 2025. 4