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

*Xinyu Wei*[1,4*], *Kangrui Cen*[4*], *Hongyang Wei*[2,4*], *Zhen Guo*[1,4], *Bairui Li*[1,4], *Zeqing Wang*[3,4], *Jinrui Zhang*[1,4], *Lei Zhang*[1,4†]

[1]The Hong Kong Polytechnic University [2]Tsinghua University [3]Sun Yat-sen University [4]OPPO Research Institute

[*]Equal contribution [†]Corresponding author

## Abstract

*In controllable image generation, synthesizing coherent and consistent images from multiple reference inputs, i.e., **Multi-Image Composition** (MICo), remains a challenging problem, partly hindered by the lack of high-quality training data. To bridge this gap, we conduct a systematic study of MICo, categorizing it into 7 representative tasks and curate a large-scale collection of high-quality source images and construct diverse MICo prompts. Leveraging powerful proprietary models, we synthesize a rich amount of balanced composite images, followed by human-in-the-loop filtering and refinement, resulting in **MICo-150K**, a comprehensive dataset for MICo with identity consistency. We further build a Decomposition-and-Recomposition (De&Re) subset, where 11K real-world complex images are decomposed into components and recomposed, enabling both real and synthetic compositions. To enable comprehensive evaluation, we construct **MICo-Bench** with 100 cases per task and 300 challenging De&Re cases, and further introduce a new metric, **Weighted-Ref-VIEScore**, specifically tailored for MICo evaluation. Finally, we fine-tune multiple models on **MICo-150K** and evaluate them on **MICo-Bench**. The results show that MICo-150K effectively equips models without MICo capability and further enhances those with existing skills. Notably, **Qwen-MICo**, fine-tuned from Qwen-Image-Edit, matches **Qwen-Image-2509** in 3-image composition while supporting arbitrary multi-image inputs beyond the latter's limitation. Our dataset and benchmark will be valuable resources for advancing MICo research.*

## 1. Introduction

With the rapid advancement of generative AI, the capability of image synthesis has expanded dramatically. Both text-to-image (T2I) [5, 9, 19, 21, 22, 29, 32, 38, 42, 45, 55, 61] and image-to-image (I2I) [4, 17, 30, 34, 39, 58, 62, 70] models can now produce visually stunning and photorealistic results. Meanwhile, these models have possessed richer world knowledge and can support an increasingly diverse set of conditioning modalities [36, 43, 71], ranging from complex textual prompts to depth maps, sketches, and reference images. Among them, generating new images following the textual intent of users while maintaining identity (ID) consistency with reference images (particularly facial ID [8, 14, 23, 31, 46, 47, 54, 59, 69]), a task often referred to as personalized generation [40, 41, 60] or in-context generation [30, 51, 63, 72], represents one of the most valuable capabilities of modern generative models. Recent works such as FLUX.Kontext [30] and Qwen-Image [61] have made remarkable progress in this direction. However, these systems only support a single-reference input, limiting the ability to integrate multiple entities into a unified and coherent composition, restricting users' freedom of creative design.

**Multi-Image Composition** (MICo) with identity consistency, which involves combining content from multiple *source images* into a coherent *target image* while faithfully adhering to the textual prompts, stands as one of the most challenging yet underexplored tasks in generative image modeling. Proprietary models like GPT-Image-1 [49], Nano-Banana [15], and Seedream 4.0 [48] have pushed this task to a new level of realism, coherence, and controllability. However, open-source community remains a significant gap in MICo compared to proprietary models, partly due to the lack of high-quality datasets tailored for this task.

ConceptBed [40] provides a large-scale dataset of ID-preserving images across varying concepts but does not include MICo data. Echo-4o [25] introduces such data only as a small subset, with limited diversity in sources and prompts. Some approaches achieve MICo without explicitly constructing composed targets [12, 41, 68]. Another line of research [35], as shown in Fig. 1, constructs training pairs by segmenting instances from full images using GroundingDINO [33] and SAM [26] as sources while originals as targets [7, 24]. However, such paradigms often yield incomplete and semantically ambiguous samples. Later methods such as UNO [63], OmniGen2 [62], and DreamO [37] enhance semantic richness by synthesizing new contexts; however, the weak generative backbones they used lead to highly homogeneous results in both style and content. Recently, DreamOmni2 [64] introduces a three-stage data generation pipeline, achieving more controllable and consistent compositions. Nevertheless, the synthesized


Figure 1. Previous MICo methods typically collect high-quality images or video frames as **target images (1)**. Using Open-Vocabulary Detectors (OVD) [33] and SAM [26], objects within targets are segmented to obtain **source images (2)**. Some methods enhance the targets by retrieving additional frames of the same subject from videos **(3)**, or enhance the sources using S2I (Subject-to-Image) or inpainting models **(4)**. Training pairs are then constructed along multiple paths: (2→1), (2→3), (4→1), and (4→3). However, the masks in **(2)** are often incomplete and semantically ambiguous; the generated images in **(4)** tend to share similar styles, content, and limited diversity due to reliance on a few fixed generative models; the frames in **(3)** originate from a small number of high-quality videos, leading to limited scene variety and a lack of imaginative or complex multi-subject scenarios.

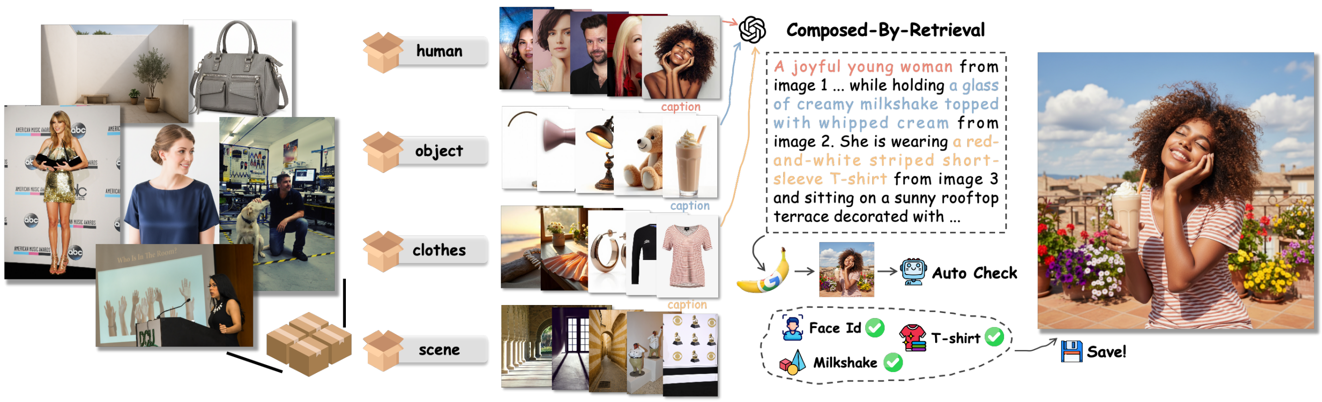
data still suffer from stylistic and contextual homogeneity.

In summary, existing MICo datasets face two main limitations: **(1)** many source or target images are generated by a few fixed T2I models, leading to homogeneous content and a noticeable quality gap compared to proprietary counterparts; **(2)** datasets based on real photos or video clips are limited in diversity, lacking imaginative scenarios, and biasing toward person-centric compositions, with insufficient coverage of object-centric or multi-subject cases.

To address these limitations, we introduce **MICo-150K**, a high-quality and comprehensive dataset specifically designed for multi-image composition tasks. We first define a taxonomy with **3 major categories**, **7 sub-tasks**, and **27 fine-grained types**, along with a realistic **Decompose and Recompose (De&Re) track**, as illustrated in Fig. 2. For each task, we design dedicated data collection and cleaning pipelines to curate high-quality source images of objects, humans, clothing, and scenes. Each image is paired with a detailed, descriptive caption aligned with its visual content. To generate composition prompts, we use GPT-4o [39] with a **Compose-by-Retrieval** strategy, which combines compatible source captions to form rich and coherent prompts for each composition type. These prompts are then fed into Nano-Banana [15], a state-of-the-art proprietary model, to synthesize high-quality composite images. Finally, a human-in-the-loop post-filtering process is applied to retain only the most accurate and visually consistent results, yielding the final MICo-150K dataset.

It is also observed that there lacks a dedicated benchmark for evaluating MICo performance. To fill this gap, we introduce **MICo-Bench**, a comprehensive benchmark comprising 1,000 carefully curated cases, 100 for each of the seven MICo sub-tasks and 300 challenging cases in the De&Re subset. Each case is rigorously validated by human annotators to ensure diverse semantics and scene coverage across source images and text prompts. For evaluation, we propose a new metric **Weighted-Ref-VIEScore**, a reliable and human-aligned metric for assessing MICo performance.

We fine-tune several T2I models, BAGEL [16], BLIP3-o [10], Lumina-DiMOO [67], OmniGen2 [62], and Qwen-Image-Edit [61] on **MICo-150K** and evaluate them using **MICo-Bench**. BLIP3-o and Lumina-DiMOO that originally lacked multi-image composition capabilities acquire strong MICo abilities after fine-tuning. BAGEL and Qwen-Image-Edit, though not explicitly trained for multi-image composition, demonstrate emergent MICo capability that is significantly enhanced after fine-tuning on MICo-150K. OmniGen2, which already possesses robust multi-image consistency, also exhibits further improvement after training. In summary, our contributions are threefold: (i) We introduce **MICo-150K**, a large-scale high-quality dataset specifically designed for multi-image composition with identity consistency, and provide comparisons between real-world and synthetic compositions.; **(ii)** We present **MICo-Bench** along with the **Weighted-Ref-VIEScore**, establishing a comprehensive and reliable evaluation framework for MICo task; **(iii)** We observe that some strongly pretrained I2I models **naturally exhibit emergent MICo capabilities**. Extensive experiments further demonstrate the robustness and generality of MICo-150K, as models with diverse architectures and initialization states consistently achieve substantial improvements after fine-tuning.




(a) High-quality open-source data are collected and cleaned through a dedicated pipeline, categorized into four groups: **human**, **object**, **clothes**, and **scene**, each with detailed captions. For each task, a diverse set of source images is randomly sampled from these categories, and a multi-image composition prompt is generated by GPT-4o using image captions under a "**Composed-by-Retrieval**" strategy. The generated prompt is then fed into Nano-Banana [15] to synthesize composite images. Each output image undergoes an automated verification process including QwenVL2.5-72B [1] and ArcFace [18] to ensure that all source images are correctly represented in the final composition before being included in the dataset.


(b) We collect a large number of high-quality single-person portraits through our data-cleaning pipeline and use **Nano-Banana** [15] to **decompose** each image into its constituent components—**scene**, **human**, **objects**, and **clothes**. Human annotators then carefully inspect all decomposed components to ensure their quality. Once all parts meet the required standards, **Nano-Banana** is used again to **recompose** them into a complete image.

Figure 2. Construction pipeline of MICo-150K. **(a)** The data construction pipeline for the **Human-Centric**, **Object-Centric**, and **HOI (Human–Object Interaction)** tasks. **(b)** The pipeline for the **DeRe (Decompose and Recompose)** task.

## 2. Related Work

**Training-free MICo Methods**. Versatile Diffusion [68], trained solely on standard I2I data, adopts a unified multi-flow diffusion framework optimized across diverse tasks, exhibiting certain MICo capabilities. $\lambda$-ECLIPSE [41] learns the CLIP latent space using image–text interleaved data and achieves MICo through a frozen diffusion inference backend without any specific supervision. MIP-Adapter [24] introduces lightweight, learnable adapters into a frozen diffusion model, enabling MICo by training a small adapter. LAMIC [12] extends single-reference T2I models in a fully training-free manner by introducing *Group Isolation Attention* and *Region-Modulated Attention*, which support layout-aware MICo without modifying the underlying MMDiT parameters or requiring additional training data.

**Trainable MICo Methods**. Subject Diffusion [35] pioneers a data construction paradigm that extracts background-free instances using GroundingDINO [33] and SAM [26] as source images, with the original images serving as targets. This strategy has provided valuable insights for building MICo datasets. MS-Diffusion [60] follows this approach but constructs semantically aligned image pairs from different frames of the same object in video clips, avoiding direct copy–paste duplication. UNO [63] enhances model capability through progressive cross-modal alignment. It follows Subject Diffusion [35] but employs a Subject-to-Image (S2I) model to generate new source images rather than using the original instances directly. DreamO [37] introduces *feature routing constraints* into the DiT training process to ensure identity consistency. It uses PuLID [23] to generate multiple portraits of the same person ID and selects different frames of the same subject from videos for body-level data. OmniGen2 [62] employs two distinct decoding pathways for text and image with unshared parameters and achieves strong MICo ability through large-scale training. Its data pipeline also leverages multi-frame videos but includes refined filtering and inpainting strategies to restore background context. XVerse [7] introduces a *T-Mod Resampler* that processes each source image and injects its representation into a per-token modulation adapter, while incorporating VAE features to preserve fine-grained details. DreamOmni2 [64] proposes a novel three-stage data construction pipeline. It first introduces *dual-branch attention mixing* in a T2I model to generate two semantically aligned images with partially shared features, then trains an extraction model on them, which is subsequently used to generate reference images.

## 3. MICo-150K Dataset

This section provides the detailed construction pipeline of the **MICo-150K**. Sec. 3.1 describes the collection of source images, while Sec. 3.2 details the construction processes of the four task categories. For clarity, the number of cases for each sub-task is summarized in Tab. 1, representative examples from each task are illustrated in Fig. A1 in **Appendix**.

### 3.1. Source Image Collection

**Object Images**. We collected object images and the corresponding captions from the Subject200k [52] dataset. To ensure quality, we employed **Qwen2.5-VL-72B** [1] for filtering, removing images containing humans, obvious blur, corruption artifacts, or ambiguous semantics. Each retained image was assigned a detailed, descriptive caption and category label. Within each category, we extracted visual features using **DINO-v3** [50] and **SigLIP2** [56], concatenated them, and applied **DBSCAN** [20] clustering to eliminate redundancy, keeping only one representative image per visual-semantical cluster. After this rigorous cleaning process, we obtained **31.5K** high-quality object images.

**Human Images**. We collected 13.7K high-quality model images from the VITON-HD [13] try-on dataset, 6K upper-body portraits from the Headshot iStockPhoto dataset [2], 3K high-resolution portraits from the Headshot Pexels v1 dataset [3], professional portrait photos from the X2I-Subject-Driven dataset [66], and 144K photos of 5,403 celebrities. We then employed **Qwen2.5-VL-72B** [1] to remove images containing multiple faces, back views, or heavily occluded faces. Each remaining image was annotated with a detailed caption describing gender, attributes, clothing, and posture. After the cleaning and annotation process, we obtained **44.6K** high-quality human portraits, ensuring maximal diversity in facial identities.

**Cloth Image**. We collected clothing images from the Subject200k [52] and the VITON-HD [13] try-on dataset. The clothing items were categorized by type (upper garments, lower garments, shoes, accessories) and gender, forming $4_{type} * 2_{gender} = 8$ subcategories in total. We used Qwen2.5-VL-72B [1] to filter out unclear or inappropriate items and generate descriptive captions for each image. The final source pool contains **17.3K** upper garments, **1.3K** pants, **428** pairs of shoes, and **7.8K** accessories.

**Scene Images**. We collected scene images from the Mulan [57], Echo-4o [25], and SUN397 [65] datasets. However, the overall quality of these scenes varies significantly. To make them suitable, we first filtered out images whose shorter side was below 512 pixels. For the remaining images, we used Qwen2.5-VL-72B [1] to perform a series of quality checks: *(1)* the scene must be open and contain no people; *(2)* it should depict a realistic, natural environment with no visible artifacts; *(3)* it should be clean and well-organized, without elements such as toilets or garbage; *(4)* it must not be overly confined and should be spacious enough to accommodate at least one person. In addition, we used GPT-4o [53] to generate extra portrait-friendly scenes. We manually collected 200 high-frequency scene descriptions from AIGC community forums used for portrait generation, assigned concise text labels to them, and then prompted GPT-4o to synthesize the corresponding scenes. After all filtering and generation, we obtained a final set of **11K** high-quality scenes suitable for the MICo tasks.

**Compose-by-Retrieval**. We employ Nano-Banana [15] to compose the source images according to task settings, producing high-quality composite results. However, directly sampling source images at random from the pools may lead to semantically incompatible combinations, for instance, pairing a male athlete with high-heeled shoes, resulting in poor generation quality. To address this, we propose the **Compose-by-Retrieval** strategy. After choosing a primary (subject) image, we sample multiple candidates for clothing, scenes, or objects from respective pools. We then provide the subject image, candidate images, and their detailed captions as the context to GPT-4o [53] to select the most semantically compatible combination for composition.

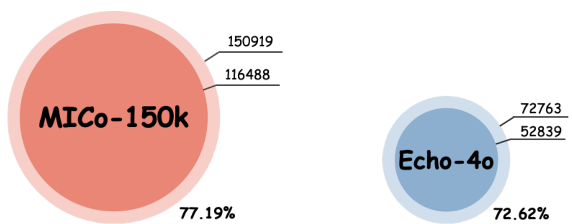
### 3.2. Construction of MICo Tasks

**Object-Centric Tasks**. We define two sub-tasks for object-centric composition. The *Object + Object* task involves combinations of one to four distinct objects, where source images are randomly sampled for composition. The *Object + Scene* task covers compositions of a single object or two objects placed within a specific scene. We adopt the proposed *Compose-by-Retrieval* strategy, which provides multiple candidate scenes for each object and adaptively selects the most semantically compatible one. Previous MICo methods typically use source image captions directly as prompts, such as "`Combine 2 images according to <Caption A> and <Caption B>.`" To promote prompt diversity and naturalness, we instead treat source captions as contextual input for GPT-4o, which generates a more coherent composition prompt. Furthermore, GPT-4o annotates the prompt with explicit token-to-source mappings, providing a foundation for future studies on latent space alignment. We then used Nano-Banana to synthesize composite images and employed Qwen2.5-VL-72B for verification. Each source image is paired with the gener-

Table 1. Statistics of the **MICo-150k** dataset. The dataset is carefully balanced to ensure diversity across task types and input images. In the **Types** column, abbreviations denote component categories: O = Object, S = Scene, P = Person, M = Man, W = Woman, and C = Cloth.

| Task | Sub-task | Types | Count |
|---|---|---|---|
| **Object Centric** | Object + Scene | 1O1S, 2O1S | 5014, 4999 |
| | Object + Object | 2O, 3O, 4O, 5O | 10007, 10012, 5001, 4998 |
| **Person Centric** | Person + Person | 1M1W, 2M1W, 2W1M<br>2M, 2W, 3M, 3W | 3001, 3003, 3066<br>2996, 3006, 2999, 2991 |
| | Person + Scene | 1P1S, 2P1S | 4986, 4994 |
| **Human Object Interaction** | Person + Objects | 1P1O, 1P2O, 2P1O, 2P2O | 5020, 4973, 5071, 5050 |
| | Person + Clothes | 1P1C, 1P2C, 1P3C, 1P4C | 7031, 7017, 6942, 6973 |
| | Person + Objects + Clothes | 1P1C1O, 1P1C2O<br>1P2C1O, 1P2C2O | 5034, 5022<br>5025, 5011 |
| **De&Re** | **Adaptive** | | 11677 |


(a) Statistics of source image counts across different categories.




(b) Semantic redundancy in text prompts. We extract text embeddings using CLIP and set a cosine similarity threshold of 0.85 to identify duplicates.

Figure 3. High-quality multi-image composition datasets that are non-segmentation-based and not generated by Flux series [30] are extremely rare; to the best of our knowledge, only Echo-4o [25] is publicly available. MICo-150K significantly surpasses it in both source image diversity and text prompt semantic variety.

ated target, and the model checks whether the objects from the source appear correctly in the final composition.

**Person-Centric Tasks**. We define two sub-tasks: *Person + Person*, which covers 7 different gender and group combinations, and *Person + Scene*, which composes 1 or 2 individuals placed within specific scenes. We find that omitting explicit gender specification in prompts degrades generation quality. Therefore, we explicitly state gender information in the *Person + Person* task. For the *Person + Scene* task, we adopt the Compose-by-Retrieval strategy to select semantically compatible scenes for each person. During quality control, objects and scenes are verified using Qwen2.5-VL-72B, while face ID consistency is evaluated via ArcFace [18] embeddings. Specifically, we extract face ID embeddings from both the source and generated images and compute the optimal matching using the Hungarian algorithm [28]. A generation is considered valid only if all matched pairs exceed a task-specific similarity threshold.

**Human Object Interaction Tasks**. We define three subtasks: *Person + Object*, *Person + Clothes*, and *Person + Object + Clothes*, each further divided into four task variants as summarized in Tab. 1(third row). During the composition phase, we explicitly specify the gender of the human source and apply the Compose-by-Retrieval Strategy to select semantically compatible objects or clothes for each case. For quality verification, we employ both Qwen2.5-VL-72B and ArcFace. The former ensures that all visual entities are correctly rendered, while the latter extracts face ID and evaluates the consistency with the source images.

**Decompose-and-Recompose (De&Re) Tasks**. This task is the most complex and comprehensive part in MICo-150K. We first curated a large collection of high-quality single-person portraits from the CC12M [6] dataset. Using Nano-Banana, we decomposed these complex, realistic photographs into separate components such as humans, clothing, objects, and scenes. Human annotators then conducted fine-grained quality checks, identifying failure cases, such as those missing object identities, or directly copy-pasted, lacking sufficient variance. Then, the annotators rewrote component-specific prompts to correctly extract the intended elements. After this human-in-the-loop refinement process, Nano-Banana was used again to recombine them into complete images. Thus, each set of components produces two versions: a real-world composition and a recomposed synthesized one, as illustrated in Fig. 2.

After the above construction process, the statistics of the dataset's source images and text prompts are shown in Fig. 3, demonstrating significantly greater diversity compared to other datasets [25] with similar functionality.

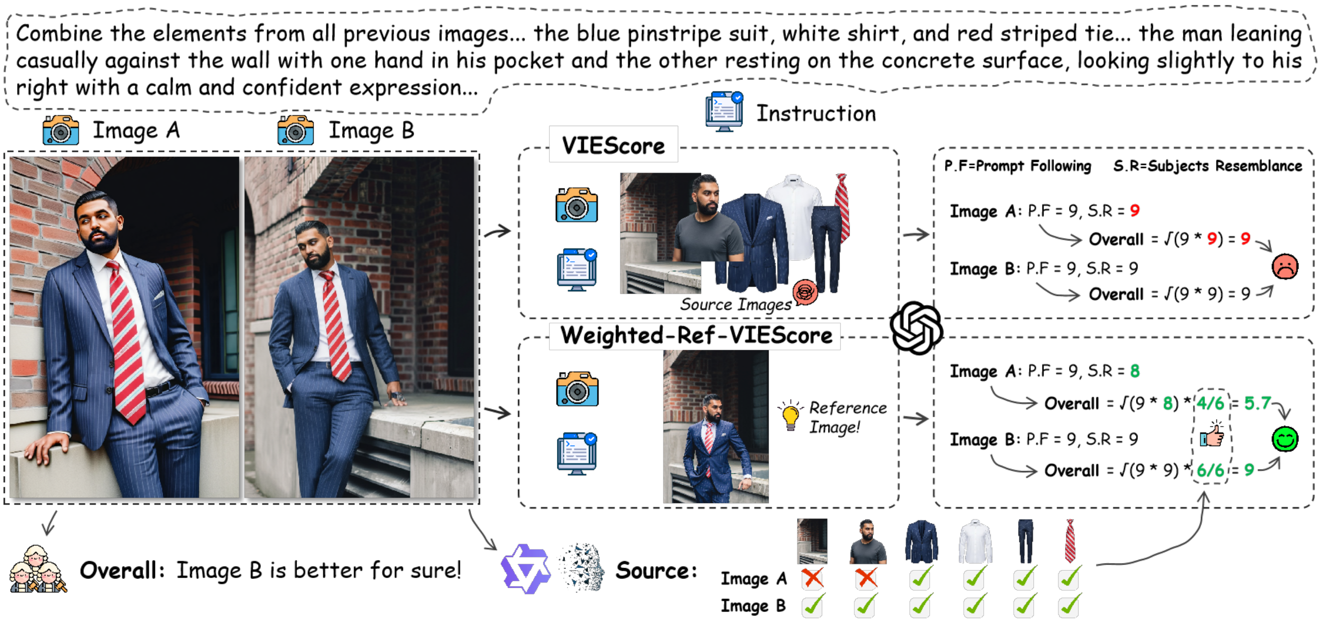



Figure 4. Traditional VIEScore requires inputting all source and generated images into the evaluator, which often leads to degraded performance as **GPT-4o's cross-image attention becomes overloaded**. This prevents the model from fully understanding each image and accurately determining whether every source appears in the target, resulting in substantial scoring errors (in this example, all the three human evaluators unanimously agreed that Image B was far superior). In contrast, MICo-Bench first assess whether each source image appears in the generated result to produce weights. Each case also includes a verified reference image that contains all sources. During evaluation, GPT-4o compares only the generated image and the reference image, enabling human-level judgment accuracy.

## 4. MICo-Bench

We first analyze why conventional image generation metrics are insufficient for MICo task (Sec. 4.1), then introduce the Weighted-Ref-VIEScore (Sec. 4.2). Finally, we describe the curation pipelien of our benchmark (Sec. 4.3).

### 4.1. Why Existing I2I Metrics Fail for MICo?

With the rapid rise of general-purpose Vision-Language Models (VLMs), evaluating generative tasks such as I2I and T2I through VLMs has become increasingly popular. VIEScore [27] offers a representative framework for VLM-based image evaluation by decomposing overall quality into two dimensions: Semantic Consistency (**SC**) and Perceptual Quality (**PQ**). The final score is computed as $SC \times PQ$. OmniContext Bench [62] follows this framework and employs GPT-4o [53] to evaluate the **SC** score of composition results in terms of Prompt Following (**PF**) and Subject Resemblance (**SR**). However, it has a key limitation: it requires all source images to be input simultaneously during evaluation. Although modern VLMs like GPT-4o [53] are highly capable, their **cross-image attention remains limited**. When too many images are provided, the model struggles to accurately perceive each image's content, leading to unreliable assessments of composition quality and consequently wrong scores, as illustrated in first row of Fig. 4.

### 4.2. Weighted-Ref-VIEScore

To overcome the limitations of existing evaluation frameworks, we introduce a weighted, reference-based evaluation paradigm tailored for multi-image composition.

**Weighting**: During each evaluation, every non-human source image is first paired with the generated result and fed into Qwen-VL2.5-72B [1] to determine whether the source image successfully appears in the generated output. For facial source images, we instead employ ArcFace [18] to verify whether the corresponding face is correctly preserved in the result. This process yields a weight for each source image, reflecting its contribution to the composition. However, relying solely on these weights may be exploited by models, for instance, **by trivially copy-pasting all sources into the output**. To prevent such bias, we incorporate a complementary reference-based mechanism.

**Reference**: For each case, we first generate a high-quality reference image using Nano-Banana [15], followed by rigorous human verification to ensure that all source elements are faithfully represented, with the necessary visual variance and semantic alignment with the given prompt. During evaluation, GPT-4o [53] compares each generated image with its reference image individually, rather than jointly processing all source images, as illustrated in Fig. 4 (Weighted-Ref-VIEScore). . This design yields a more ac-

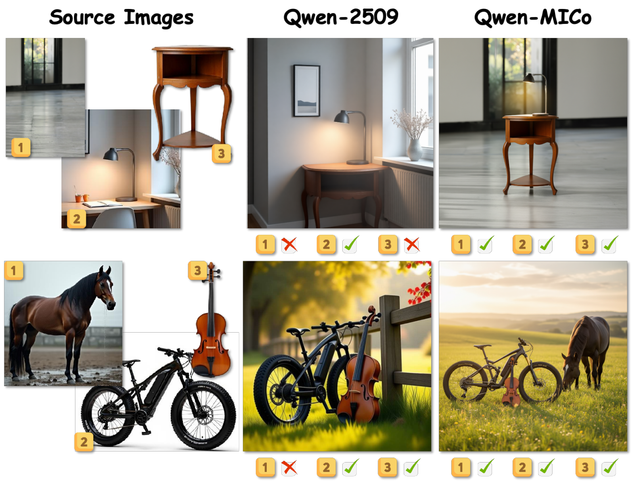



Figure 5. Qwen-Image-2509 is trained on a massive-scale dataset and supports only three-image inputs, while Qwen-MICo is trained solely on MICo-150K and supports arbitrary multi-image inputs, producing images with higher aesthetic quality.

curate and human-aligned **SC** score. The **PQ** score is then computed solely based on the generated image. Finally, the overall score is defined as $W \times \sqrt{SR \times PF} \times PQ$ where $W$ denotes the source-specific weighting factor. Extensive user studies demonstrate that our metric aligns with human preference significantly better than existing alternatives.

### 4.3. Benchmark Curation

We construct MICo-Bench by selecting 200, 200, and 300 representative cases from the Object-Centric, Person-Centric, and Human–Object Interaction tasks of MICo-150K (as shown in Tab. 1), respectively—100 cases from each sub-task, and an additional 300 cases from the De&Re subset, resulting in 1,000 high-quality benchmark cases. To ensure annotation reliability and fairness, we invited three independent reviewers who were not involved in this project to perform the selection. These reviewers were unaware of the distinctions between the general composition tasks and the De&Re task. Each case was evaluated from two perspectives: (1) whether all objects from the source images were correctly preserved in the generated result, and (2) whether the final image aligned with the textual instruction. Only the cases unanimously judged as satisfactory by all three reviewers were retained for the final benchmark.

## 5. Experiments

### 5.1. Experiments Setting

To validate the effectiveness of our dataset, we train five open-source models: BAGEL [16], OmniGen2 [62], Luminia-DiMOO [67], BLIP3-o [10], Qwen-Image-Edit [61] on MICo-150K, and evaluate their performance on MICo-Bench. For BAGEL, we freeze the understanding branch and VAE while fine-tuning the generation branch parameters. For BLIP3-o, we build upon the BLIP3o-Next-Edit [11] version and set the maximum sequence length to 5120, with sequences right-padded or truncated as needed. For OmniGen2, we simply perform full fine-tuning. For Luminia-DiMOO, we set the input maximum sequence length to 6144, and resize all input images to $512 \times 512$. For Qwen-Image-Edit, we use our custom training pipeline. The encoder and VAE are frozen, while all parameters of the MMDiT module are fine-tuned. We further conducted an ablation study on the De&Re task, comparing three training configurations: without De&Re, with real De&Re target images, and with synthetic De&Re target images.

### 5.2. Results and Discussions

The experimental results in Tab. 2 show that models with diverse architectures and initial capabilities all exhibit substantial improvements in multi-image composition after fine-tuning on MICo-150K. Notably, the fine-tuned Qwen-MICo (derived from Qwen-Image-Edit) achieves performance close to that of closed-source models, further demonstrating the effectiveness of our dataset.

**Open-Source Models**. As shown in Fig. 6, BLIP3-o and DiMOO, which originally lack any MICo ability, exhibit clear gains after fine-tuning on MICo-150K. BAGEL and Qwen-Image-Edit, benefiting from large-scale multi-task pretraining, already display *emergent MICo capabilities*, which are further reinforced after fine-tuning. The fine-tuned Qwen-MICo achieves 3-image composition performance comparable to Qwen-Image-2509 [44], as shown in Fig. 5, despite the latter being trained on a dataset hundreds of times larger. Moreover, Qwen-MICo supports arbitrary multi-image inputs, whereas Qwen-Image-2509 is limited to three. OmniGen2, pretrained on MICo-related tasks, already exhibits compositional ability, especially in maintaining facial identity, and further benefits from MICo-150K.

**Closed-Source Models**. Overall, Nano-Banana achieves higher quantitative scores, yet GPT-4o exhibits superior robustness. Regardless of the input configuration, GPT-4o consistently produces coherent and semantically appropriate compositions, avoiding issues such as incomplete bodies, total identity loss, or direct "copy–paste" artifacts that occasionally appear in Nano-Banana's outputs. Representative cases are illustrated in Fig. A3 of the **Appendix**.

**Emergent Capability.** It is worth noting that BAGEL and Qwen-Image-Edit were not trained on any multi-image composition data. However, *when multiple source image tokens are simply concatenated and fed into the model, they exhibit emergent MICo ability*. After simple SFT on MICo-150K, these abilities are significantly enhanced. Moreover, even without using data from the De&Re task, where no

Table 2. Performance comparison on MICo-Bench across different open-source and closed-source model: "base" denotes the original model; "w/o" indicates fine-tuning without the De&Re task; "real" and "synth" correspond to fine-tuning with real and synthetic compositions from the De&Re task, respectively. The best performance of each model under each task is highlighted in **bold**.

| Model | Overall | | | | Object | | | | Person | | | | HOI | | | | De&Re | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | base | w/o | real | synth | base | w/o | real | synth | base | w/o | real | synth | base | w/o | real | synth | base | w/o | real | synth |
| ***Open-Source Models*** | | | | | | | | | | | | | | | | | | | | |
| BLIP3-o | 2.2 | 42.2 | **43.2** | 43.0 | 3.2 | 36.7 | **36.7** | 36.6 | 2.1 | 41.2 | **41.3** | 41.3 | 2.1 | 45.6 | **46.7** | 46.6 | 1.2 | 42.1 | **45.2** | 45.1 |
| DiMOO | 4.3 | 32.3 | **34.2** | 33.9 | 3.8 | 33.8 | **33.8** | 33.6 | 1.3 | 31.6 | **31.7** | 31.6 | 6.5 | 39.5 | 40.1 | **40.2** | 1.8 | 23.9 | **27.0** | 26.7 |
| BAGEL | 33.3 | 42.6 | **44.3** | 44.1 | 31.4 | 39.3 | **39.3** | 39.2 | 32.4 | 44.6 | **44.8** | 44.6 | 26.9 | 38.0 | **38.9** | 38.9 | 41.2 | 48.1 | **50.9** | 50.6 |
| Qwen-Image | 38.5 | 56.4 | **58.2** | 58.1 | 35.4 | 53.0 | **53.1** | 53.0 | 38.8 | 56.4 | **56.6** | 56.5 | 36.9 | 56.5 | **56.9** | 56.8 | 41.4 | 58.2 | **60.7** | 60.6 |
| OmniGen2 | 41.0 | 50.6 | **51.2** | 50.7 | 39.0 | 42.1 | **42.2** | 42.1 | 33.1 | 50.1 | **50.4** | 50.2 | 39.5 | 54.4 | **54.8** | 54.8 | 46.9 | 51.9 | **52.4** | 52.3 |
| ***Closed-Source Models*** | | | | | | | | | | | | | | | | | | | | |
| GPT-4o | **59.6** | – | – | – | **59.1** | – | – | – | **53.8** | – | – | – | **60.1** | – | – | – | **61.3** | – | – | – |
| Nano-Banana | **60.3** | – | – | – | **58.6** | – | – | – | **59.4** | – | – | – | **62.1** | – | – | – | **60.4** | – | – | – |


Figure 6. The leftmost displays the source and reference images. The first row shows model outputs before fine-tuning, the second row presents outputs after fine-tuning. The Weighted-Ref-VIEScore for each generated result is annotated in the corner. MICo-150K demonstrates strong robustness: **BLIP-3o** and **Lumina-DiMOO** acquire MICo capability from scratch; the emergent MICo abilities of **BAGEL** and **Qwen-Image** are significantly strengthened; **OmniGen2** achieves further improvement on top of its already strong performance.

other training samples contain combined human, object, clothing, and scene all together, all models still demonstrate emergent performance on the MICo-Bench De&Re subset.

**Synthetic Data vs. Real Data**. Each group of source images in the **De&Re** task is paired with a real and a synthetic target . As shown in Tab. 2, models trained on synthetic targets achieve nearly identical performance to those trained on real ones, showing that carefully curated synthetic data can serve as an effective substitute in MICo training.

## 6. Conclusion

We introduced the **M**ulti-**I**mage **Co**mposition (**MICo**) task, along with **MICo-150K**, the first large-scale high-quality dataset tailored for this task, encompassing three tasks, seven subtasks, and 27 composition types, along with a **De&Re** subset providing real–synthetic pairs for controlled comparison, totaling 150K MICo cases with varying numbers of input images. We further proposed **MICo-Bench** and **Weighted-Ref-VIEScore**, providing a reliable framework for evaluating MICo performance. Extensive experiments reveal interesting insights and further show that training on MICo-150K consistently boosts MICo performance across different models with diverse initial strengths. Notably, Qwen-MICo fine-tuned on MICo-150K achieves 3-image composition performance comparable to Qwen-Image-2509, despite the latter being trained on a dataset hundreds of times larger and restricted to 3-image inputs, while Qwen-MICo supports arbitrary inputs.

**Limitations**. MICo-150K covers a wide variety of input combinations, but ultra-complex scenes, such as rare objects, highly dynamic actions or 5+ distinct individuals simultaneously, could be further expanded in future work.

## 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. 3, 4, 6

[2] BKM1804. headshot_istockphoto: Headshot image dataset from istockphoto. https://huggingface.co/datasets/BKM1804/headshot_istockphoto, 2025. Accessed: 2025-11-05, 6 000 images. 4

[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. 4

[4] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 1

[5] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, Yang Li, Zhenxi Li, Zhimin Li, Jiaxin Lin, Linus, Lucaz Liu, Shu Liu, Songtao Liu, Yu Liu, Yuhong Liu, Yanxin Long, Fanbin Lu, Qinglin Lu, Yuyang Peng, Yuanbo Peng, Xiangwei Shen, Yixuan Shi, Jiale Tao, Yangyu Tao, Qi Tian, Pengfei Wan, Chunyu Wang, Kai Wang, Lei Wang, Linqing Wang, Lucas Wang, Qixun Wang, Weiyan Wang, Hao Wen, Bing Wu, Jianbing Wu, Yue Wu, Senhao Xie, Fang Yang, Miles Yang, Xiaofeng Yang, Xuan Yang, Zhantao Yang, Jingmiao Yu, Zheng Yuan, Chao Zhang, Jian-Wei Zhang, Peizhen Zhang, Shi-Xue Zhang, Tao Zhang, Weigang Zhang, Yepeng Zhang, Yingfang Zhang, Zihao Zhang, Zijian Zhang, Penghao Zhao, Zhiyuan Zhao, Xuefei Zhe, Jianchen Zhu, and Zhao Zhong. Hunyuanimage 3.0 technical report, 2025. 1

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts, 2021. 5

[7] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation, 2025. 1, 3

[8] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation, 2024. 1

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 1

[10] 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. 2, 7

[11] Jiuhai Chen, Le Xue, Zhiyang Xu, Xichen Pan, Shusheng Yang, Can Qin, An Yan, Honglu Zhou, Zeyuan Chen, Lifu Huang, Tianyi Zhou, Junnan Li, Silvio Savarese, Caiming Xiong, and Ran Xu. Blip3o-next: Next frontier of native image generation, 2025. 7

[12] Yuzhuo Chen, Zehua Ma, Jianhua Wang, Kai Kang, Shunyu Yao, and Weiming Zhang. Lamic: Layout-aware multi-image composition via scalability of multimodal diffusion transformer, 2025. 1, 3

[13] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization, 2021. 4

[14] Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 950–959, 2024. 1

[15] 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, 2, 3, 4, 6

[16] 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. 2, 7

[17] 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, 2025. 1

[18] 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. 3, 5, 6

[19] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 1

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining*, 1996. 4

[21] Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xuanda Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 3.0 technical report, 2025. 1

[22] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Xin Xia, Xuefeng Xiao, Linjie Yang, Zhonghua Zhai, Xinyu Zhang, Qi Zhang, Yuwei Zhang, Shijia Zhao, Jianchao Yang, and Weilin Huang. Seedream 2.0: A native chinese-english bilingual image generation foundation model, 2025. 1

[23] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment, 2024. 1, 3

[24] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation, 2024. 1, 3

[25] 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, 4, 5

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1, 2, 3

[27] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation, 2024. 6

[28] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. 5

[29] Black Forest Labs. Flux.1 – official inference repository for flux models. https://github.com/black-forest-labs/flux, 2024. Accessed: 2025-10-18. 1

[30] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 5

[31] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 1

[32] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang, Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang, Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu, Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Mingtao Chen, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu, Wei Liu, Di Wang, Yong Yang, Jie Jiang, and Qinglin Lu. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding, 2024. 1

[33] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. 1, 2, 3

[34] Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing, 2025. 1

[35] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion:open domain personalized text-to-image generation without test-time fine-tuning, 2024. 1, 3

[36] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models, 2023. 1

[37] Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, Mengtian Li, Mingcong Liu, Yi Zhang, Shaojin Wu, Songtao Zhao, Jian Zhang, Qian He, and Xinglong Wu. Dreamo: A unified framework for image customization, 2025. 1, 3

[38] OpenAI. Dall·e 3. https://openai.com/index/dall-e-3/, 2023. Accessed: 2025-10-14. 1

[39] OpenAI. Gpt-4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025. Accessed: 2025-10-18. 1, 2

[40] Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Conceptbed: Evaluating concept learning abilities of text-to-image diffusion models, 2024. 1

[41] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. $\lambda$-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space, 2024. 1, 3

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 1

[43] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, Stefano Ermon, Yun Fu, and Ran Xu. Unicontrol: A unified diffusion model for controllable visual generation in the wild, 2023. 1

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

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023. 1

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2024. 1

[48] Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, Xiaowen Jian, Huafeng Kuang, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, Wei Liu, Yanzuo Lu, Zhengxiong Luo, Tongtong Ou, Guang Shi, Yichun Shi, Shiqi Sun, Yu Tian, Zhi Tian, Peng Wang, Rui Wang, Xun Wang, Ye Wang, Guofeng Wu, Jie Wu, Wenxu Wu, Yonghui Wu, Xin Xia, Xuefeng Xiao, Shuang Xu, Xin Yan, Ceyuan Yang, Jianchao Yang, Zhonghua Zhai, Chenlin Zhang, Heng Zhang, Qi Zhang, Xinyu Zhang, Yuwei Zhang, Shijia Zhao, Wenliang Zhao, and Wenjia Zhu. Seedream 4.0: Toward next-generation multimodal image generation, 2025. 1

[49] Jerry Sima, Eric Cheng, William Fedus, Miles Brundage, Mark Chen, Iason Gabriel, Sandhini Agarwal, Lilian Weng, et al. Addendum to gpt-4o system card: Native image capabilities. https://www.semanticscholar.org/paper/0c9b799e0dde7dcbe42f8dc61b242a0106739eba, 2024. Accessed: 2025-10-18. 1

[50] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 4

[51] Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit, 2025. 1

[52] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. 4

[53] OpenAI Team. Gpt-4o system card, 2024. 4, 6

[54] Darian Tomašević, Fadi Boutros, Chenhao Lin, Naser Damer, Vitomir Štruc, and Peter Peer. Id-booth: Identity-consistent face generation with diffusion models. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*, page 1–10. IEEE, 2025. 1

[55] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. grpo. *arXiv preprint arXiv:2505.17017*, 2025. 1

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 4

[57] 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. 4

[58] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. Seededit 3.0: Fast and high-quality generative image editing, 2025. 1

[59] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. 1

[60] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance, 2025. 1, 3

[61] 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. 1, 2, 7

[62] 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. 1, 2, 3, 6, 7

[63] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation, 2025. 1, 3

[64] Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. Dreamomni2: Multimodal instruction-based editing and generation, 2025. 1, 4

[65] 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. 4

[66] 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. 4

[67] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, Jinbin 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. 2, 7

[68] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model, 2024. 1, 3

[69] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. 1

[70] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing, 2024. 1

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[72] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer, 2025. 1