MIBE: MULTI-SUBJECT INTERACTION BENCHMARK AND EVALUATOR FOR PERSONALIZED IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-subject personalized image generation requires the precise rendering of all requested reference identities and their specified interactions based on a guiding prompt. However, state-of-the-art models still struggle with this process, frequently omitting subjects, failing to preserve reference appearances, or misattributing interactions. Furthermore, existing metrics designed primarily for single-subject fidelity cannot reliably capture these errors, suffering severe degradation in ranking separability and failing to align with human preference as the subject count increases. To address this gap, we introduce Multi-subject Interaction Benchmark and Evaluator (MIBE), a unified framework comprising a Multi-subject Interaction Benchmark (MIB) and a Multi-subject Interaction Evaluator (MIE). MIB systematically covers diverse relation types and scene complexities through a decoupled data regime. This consists of a 60K-pair VLM-labeled Silver Set for scalable metric training and a 4K-pair double-blind Human Evaluation Gold Set covering a diverse range of state-of-the-art generators, with the Silver Set reaching 95.1% cross-VLM preference agreement. To demonstrate the utility of this benchmark, we present MIE, a lightweight, reference-conditioned evaluator trained exclusively on the Silver Set with a dual-head ranking and diagnosis objective. MIE exhibits strong cross-generator generalization on the Gold Set, achieving 0.922 overall pairwise accuracy against human preference, including 0.982 on seen generators and 0.884 on unseen generators. By outperforming a broad spectrum of baseline metrics, including CLIP and DINO variants, MIE demonstrates that diagnostic supervision can preserve ranking separability and human alignment where traditional evaluators collapse. Ultimately, by providing data with the diagnostic and preference labels required to train robust evaluators, MIBE serves as a comprehensive framework to benchmark multi-subject image generation and guide future model alignment.

1 INTRODUCTION

State-of-the-art personalized image generators can faithfully render a single reference identity, but as soon as a second subject is introduced, let alone four or eight, the generated image often contains the right scene, the right colors, and the wrong subjects. A model conditioned on reference photos of a specific child and a specific dog, asked to depict the child playing with the dog, may instead produce a generic child with a different dog, two children and no dog, or the right child standing beside an empty patch of grass. These are not merely aesthetic flaws; they are *binding* failures, and they are largely invisible to current metrics.

We call this the *binding* problem: every requested reference subject must (1) appear in the generated image (*Existence*), (2) preserve its reference appearance without leakage from other subjects (*Appearance*), and (3) participate in the prompt-specified role, relation, or object assignment (*Interaction*). These dimensions are coupled: a missing subject often causes the model to redistribute its visual features onto surviving subjects, making appearance corruption a downstream consequence of existence failure (Appendix A). The problem is especially fragile in contact-rich and occlusion-heavy scenes, where local evidence is ambiguous. Prior stress-test evaluation has documented this "illusion of scalability," showing identity collapse in dense scenes while global CLIP-based scores

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

remain misleadingly stable Chen et al. (2026); MIBE complements that curated identity-collapse stress test with human-grounded labels, factorized binding stress, and a learned evaluator.

Existing automatic metrics are fragmented. CLIP- and DINO-based reference similarity captures appearance overlap but cannot verify which subject participates in which interaction. General preference models such as PickScore Kirstain et al. (2023), ImageReward Xu et al. (2023), and HPS Wu et al. (2023) can favor visually polished but compositionally broken images. As Figure 1 shows, as subject count grows from 2 to 8, the pairwise agreement of standard metrics with double-blind human judgment drops toward random choice; on MIB-Gold, HPS v2.1 reaches only 0.520 agreement and PickScore falls slightly below random at 0.486. Existing benchmarks are similarly underequipped, focusing on single-subject fidelity Ruiz et al. (2023); Peng et al. (2025), text-only composition Huang et al. (2023), or broad multi-subject capability Chen et al. (2025); Wang et al. (2026), without factorizing the conditions that induce binding failures. Concretely, existing benchmarks largely lack human pairwise gold labels for reliable, cross-generator ranking, and they report a single aggregate score that cannot isolate which failure mode, such as appearance leakage or interaction misassignment, is responsible for a low score. Binding is therefore both a generation and a measurement bottleneck.

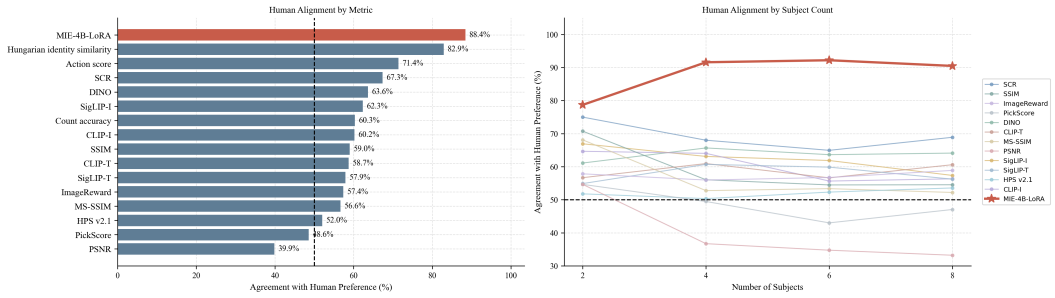


Figure 1: Similarity- and preference-based metrics align weakly with human preference, and several degrade further as subject count grows; MIE retains higher human alignment. Pairwise agreement with double-blind human preference on MIB-Gold. Left: agreement on the unseen-generator split. Right: agreement by subject count ($N \in \{2, 4, 6, 8\}$).

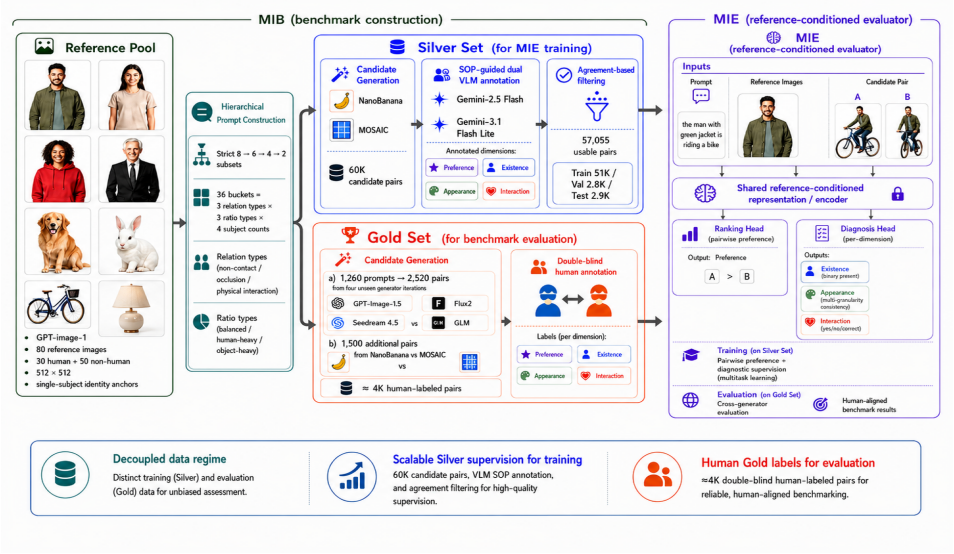


Figure 2: Overview of MIBE. MIB constructs a controlled benchmark through reference pooling, hierarchical prompt construction, Silver Set supervision, and Gold Set human evaluation. MIE is trained on Silver labels and evaluated against double-blind human Gold labels for cross-generator, human-aligned assessment.

We address this gap with **MIBE**, a unified framework that pairs a controlled **Multi-subject Interaction Benchmark (MIB)** with a learned **Multi-subject Interaction Evaluator (MIE)**. As summarized in Figure 2, MIBE decouples benchmark construction from evaluator learning: MIB defines controlled prompt-reference tasks and silver/gold supervision, while MIE uses this supervision to learn human-aligned ranking and diagnostic evaluation.

In summary, this paper makes three contributions: **1)** We introduce **MIB**, a controlled benchmark for reference-conditioned multi-subject generation that factorizes subject count, human/object composition, and relation type, with a 60K dual-VLM-consensus silver set and a 4,020-pair double-blind human-labeled gold set covering six state-of-the-art generators. **2)** We propose **MIE**, a reference-conditioned evaluator that jointly learns pairwise ranking and structured diagnostics over *Existence*, *Appearance*, and *Interaction*, providing both human-aligned preference prediction and interpretable failure attribution. **3)** Through cross-generator meta-evaluation on MIB-Gold, we reveal systematic failure patterns, failure rates that scale with subject count, coupled Existence–Appearance failures, and interaction-induced subject deformation, providing actionable targets for future model alignment.

2 Related Work

Personalized and Multi-Subject Generation Personalized image generation has evolved from early optimization-based methods like DreamBooth Ruiz et al. (2023) and Textual Inversion Gal et al. (2022) to efficient encoder-based models such as IP-Adapter Ye et al. (2023) and PhotoMaker Li et al. (2024). While effective for single subjects, these methods struggle with binding in multi-subject scenarios, ensuring each reference identity is correctly assigned to its specific role and attributes. Recent works like FastComposer Xiao et al. (2025), MS-Diffusion Wang et al. (2025), and Interact-Custom Xu et al. (2025) address this through localized attention and identity decoupling. However, a significant measurement gap remains: existing methods are rarely tested in contact-rich or occlusion-heavy regimes where subjects overlap, leading to frequent failures in physical grounding and identity leakage that current metrics fail to capture.

Benchmark Evolution While compositional benchmarks like T2I-CompBench Huang et al. (2023) evaluate text-to-image alignment, they lack the reference conditioning necessary to measure identity preservation. Recent efforts such as XVerseBench Chen et al. (2025), PSRBench Wang et al. (2026), and MultiHuman-Testbench Borse et al. (2025) have introduced multi-subject evaluations, yet they primarily focus on subject count or spatial positioning. Unlike MultiBind Tian et al. (2026), which explores attribute-level misbinding, MIB factorizes complex physical interactions and occlusions. It serves as both a diagnostic testbed and a training substrate, decoupling scalable "silver" supervision from human "gold" evaluation to provide a more granular view of how models handle locally ambiguous visual evidence. A detailed comparison with prior benchmarks is provided in Appendix B.

Metrics and Scalable Evaluators Standard metrics often provide incomplete assessments of multi-subject quality: CLIP-based scores lack instance-specific grounding, while reward models like ImageReward Xu et al. (2023) often prioritize aesthetics over logical binding. Although Vision-Language Models (VLMs) are increasingly used as automated judges, they remain sensitive to prompting and poorly calibrated for absolute scoring Kumar et al. (2026). MIBE addresses these limitations by adopting a pairwise ranking format and a latent scalar utility model, rather than absolute numeric ratings. By combining an "errors-first" VLM consensus for scalable MIB-Silver labels with human-verified MIB-Gold data, MIBE unifies preference learning with structured diagnostic supervision for more robust, reference-grounded evaluation.

3 Methods

3.1 MIB Construction

MIB is the data substrate of MIBE, consisting of controlled prompt-reference tasks, paired generated candidates, and preference/diagnostic labels. Each task specifies a guiding prompt, a set of reference subjects, and two candidate generations produced under the same prompt-reference condition. This decoupled design separates reference construction, prompt-task design, candidate generation, and

label acquisition, allowing subject identity, scene complexity, relation type, generator source, and annotation source to be controlled independently.

3.1.1 Prompt Construction

We construct 60K prompt-reference tasks from 15K hierarchical Level-8 seeds. Each seed starts from an eight-subject scene and is reduced into Level-6, Level-4, and Level-2 variants by removing entities together with their associated relations, as illustrated in Figure 3. This strict subset construction keeps scene semantics approximately invariant across levels, so performance degradation can be more directly attributed to increasing entity-relation density. The prompt space is factorized into 36 buckets from 4 subject counts, 3 human/non-human ratio types, and 3 relation types. MIB-Silver's 60K prompts split 15,000 each across the four subject counts, 20,004/20,004/19,992 across non-contact/occlusion/physical-interaction relations, and 20,000/20,000/20,000 across balanced/human-heavy/object-heavy ratios, with each of the 36 subject-count $\times$ relation $\times$ ratio buckets holding 1,666–1,667 prompts; MIB-Gold is stratified the same way, with 1,260 prompts split exactly 315 per subject count and balanced over relation and ratio type. Relation types include non-contact relations, occlusion relations, and physical interactions. When relations overlap, we assign the task to the most restrictive category using the priority physical interaction $>$ occlusion $>$ non-contact. To keep role assignment interpretable, each entity participates in at most one core relation.

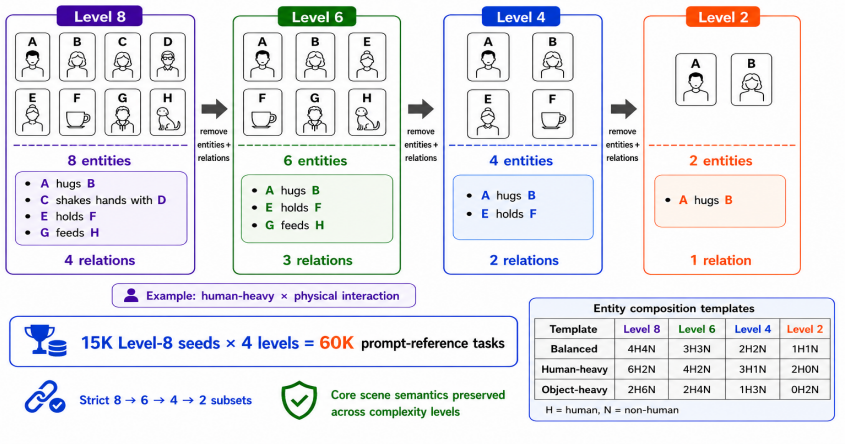


Figure 3: Hierarchical prompt construction and entity-composition templates. Each Level-8 seed is reduced into strict Level-6/4/2 subsets by removing entities together with their associated relations, preserving core scene semantics while varying entity-relation density. The example illustrates a human-heavy physical-interaction bucket, while the table summarizes the human/non-human composition templates used across balanced, human-heavy, and object-heavy ratio types.

All prompts use a white-studio setting and a global realistic-scale constraint to reduce background, lighting, layout, and scale confounders. We apply rule-based parsing and a lightweight LLM validator to verify: (i) assigned entities are explicitly mentioned, (ii) unassigned entities are excluded, (iii) human/non-human counts match the target bucket, (iv) lower-complexity prompts remain strict subsets of higher-complexity prompts, (v) duplicate prompts are removed, (vi) no entity is reused as the subject of multiple independent relations, and (vii) interaction-specific keywords are excluded from non-contact prompts.

3.1.2 Reference and Candidate Image Generation

MIB uses a fixed pool of 80 synthetic reference subjects, including 30 diverse human identities (varying in age, gender, ethnicity, and attire) and 50 non-human categories. To introduce deliberate difficulty, many subjects share similar attributes (e.g., casual wear) to prevent models from relying on trivial visual shortcuts. The detailed demographic and categorical distribution of this reference pool is provided in Appendix C.

To construct the MIB-Silver Set, we leverage two image generators, Nano Banana and MOSAIC, to generate a candidate image pair for every single one of the 60K prompt-reference configurations. This matchup was selected to maximize failure diversity observed in preliminary inspection, including subject omission, identity loss, appearance leakage, and relation hallucination. These complementary errors yield informative pairwise supervision across Existence, Appearance, and Interaction. The reference generator is also model-disjoint from the two candidate generators, reducing generator-specific leakage in scalable supervision.

MIB-Gold contains additional 4,020 pairs: 2,520 pairs from two fixed cross-model matchups among GPT-Image-1.5, Flux2, Seedream 4.5, and GLM, plus 1,500 Nano Banana-versus-MOSAIC pairs for seen-generator evaluation. Candidate order is randomized, and invalid generations are removed before annotation. Dataset statistics are summarized in Table 1.

Attribute	**Silver Set**	**Gold Set**
Purpose	MIE training	Human benchmark evaluation
Scale	60,000 candidate pairs; 56,909 usable pairs	4,020 human-labeled pairs
Prompt coverage	Hierarchical prompts with strict high-to-low complexity subsets, approximately uniformly distributed across relation type, human/object ratio, and subject count.	
Candidate generators	Nano Banana, MOSAIC	GPT-Image-1.5, Flux2, Seedream 4.5, GLM; plus Nano Banana/MOSAIC for seen-generator evaluation
Annotation source	SOP-guided dual VLM consensus	Double-blind human annotation
Annotators	Gemini-2.5-Flash and Gemini-3.1-Flash-Lite	Two independent human annotators per pair
Label types	Pairwise preference; Existence; Appearance; Interaction	Pairwise preference; Existence; Appearance; Interaction
Evaluation role	Train and Val	Seen-generator: 1,500 pairs; unseen-generator: 2,520 pairs

Table 1: Dataset statistics for MIB. The Silver Set provides scalable SOP-guided VLM supervision for training MIE, while the Gold Set is fully human-labeled under a double-blind protocol.

3.1.3 Label Construction

MIB provides two complementary label types: pairwise preference labels for global ranking and diagnostic labels for failure localization. Pairwise preference asks which candidate is better overall under the same prompt-reference condition, avoiding the calibration difficulty of absolute scores while preserving relative severity information. Diagnostic labels decompose multi-subject generation quality into three dimensions: 1) **Existence**: whether all requested reference subjects are present without omission, duplication, or severe unidentifiable collapse. 2) **Appearance**: whether generated subjects preserve the structure and local appearance of their references without severe distortion or cross-subject feature bleeding. 3) **Interaction**: whether subject-to-subject relations, actions, and object assignments align with the prompt. For diagnostic labels, annotator disagreements are averaged as soft scores, while pairwise preferences require strict consensus.

MIB-Silver annotation with SOP-guided VLM judges. The SOP decomposes each comparison into Existence, Appearance, and Interaction checks and requires candidate-specific flaw logs before the final preference decision, grounding judgments in concrete failures rather than holistic aesthetics. We independently query gemini-2.5-flash and gemini-3.1-flash-lite-preview with the same SOP, retain pairs with matching pairwise preferences, and average their per-dimension diagnostic outputs as soft supervision. This dual-judge configuration was selected via a 100-pair pilot human-alignment study: each judge alone reaches 98.9% pair-level alignment with human winners, and their strict-consensus combination retains 86.8% of pilot pairs at human alignment. Across 59,852 valid overlapping Silver comparisons, the judges reach 95.1% preference agreement, yielding 56,909 usable Silver pairs after agreement-based filtering. The full SOP appears in Appendix D.

MIB-Gold annotation with double-blind human evaluation. Each Gold pair is independently reviewed by two annotators under a double-blind protocol, providing both pairwise preference and Existence/Appearance/Interaction labels. After preference-consistency filtering, both Gold subsets

retain over 90% of pairs, indicating stable human judgments while preserving enough difficulty to expose non-trivial generator and evaluator disagreement.

3.2 Evaluator Modeling

The primary goal of MIBE is to establish a rigorous evaluation framework for multi-subject personalized image generation. To demonstrate the reusability of our benchmark and provide an immediate, out-of-the-box tool for the community, we instantiate the **Multi-subject Interaction Evaluator (MIE)**. Trained exclusively on MIB-Silver and assessed on MIB-Gold, MIE bridges the gap between human judgment and automated metrics.

Input-Output Formulation. For each task, let p denote the prompt, $R = \{r_1, \ldots, r_N\}$ denote the set of reference images, and let y denote a candidate generation. MIE is a reference-conditioned multimodal evaluator that maps the triplet (R, p, y) to two types of outputs: 1) A scalar quality score $s_\theta(R, p, y) \in \mathbb{R}$ used for global pairwise ranking. 2) A three-dimensional diagnostic prediction vector of logits $\hat{d}_\theta(R, p, y) = (\hat{d}_{\text{exist}}, \hat{d}_{\text{app}}, \hat{d}_{\text{inter}})$, corresponding to *Existence*, *Appearance*, and *Interaction*.

Prompt Construction. MIE operates as a scoring head over a fixed multimodal encoding rather than an LLM generating natural language. For a task with k reference subjects, the input is a single-turn message containing k reference-image placeholders, one candidate-image placeholder, and a static text block: *"You are evaluating a multi-subject personalization result. The first images are subject references. The last image is the generated candidate. Prompt: {prompt}"*. This is processed via the backbone's chat template with `add_generation_prompt=False`. The template design is fixed along three principles: 1) *Static and universal*: The identical instruction is used across all tasks and backbones, eliminating per-task prompt engineering or few-shot exemplars. 2) *Training-inference symmetry*: A unified code module builds prompts for both phases, preventing train/test mismatch. 3) *Agnostic and zero-rubric*: The prompt never names diagnostic facets (e.g., Existence). Judgments are learned entirely within the prediction heads rather than induced via an in-context LLM rubric.

Read-Out Mechanics. For a candidate pair (A, B) sharing references R and prompt p, their shared embeddings are pre-encoded and cached for efficiency. A and B are then processed in independent forward passes without autoregressive generation. For each candidate $X \in \{A, B\}$, the last non-padding token's hidden state h_X is extracted from the final layer. The two heads directly read out the scalar score $s_\theta(R, p, X)$ and logits $\hat{d}_\theta(R, p, X)$ from h_X. The pairwise decision is computed as $\hat{y}_{(A>B)} = \text{sign}(s_\theta(R, p, A) - s_\theta(R, p, B))$, and per-facet probabilities as $\sigma(\hat{d}_\theta(R, p, X))$.

Dual-Head Design. A purely diagnostic model with binary labels cannot capture relative failure severity (e.g., missing one subject vs. missing all). Conversely, a purely scalar reward model acts as an opaque black box, offering no actionable feedback on the nature of the failure. MIE solves this via a dual-head architecture. The **ranking head** provides a continuous global signal for pairwise comparison, which is essential for leaderboard benchmarking. Simultaneously, the **diagnostic head** explicitly attributes errors to Existence, Appearance, or Interaction, providing interpretable feedback for model developers. By optimizing these jointly ($\mathcal{L} = \alpha\mathcal{L}_{\text{rank}} + \beta\mathcal{L}_{\text{diag}}$), the scalar score is forced to ground itself in concrete binding failures rather than superficial aesthetics. This design mirrors the human annotation procedure directly, preserving both ranking separability and interpretability.

Lightweight Tuning. To ensure that the evaluator remains practical and highly reusable, we adopt a parameter-efficient fine-tuning strategy (LoRA and layer-only updates) over off-the-shelf vision-language backbones (e.g., Qwen3.5-VLM). This design choice proves the exceptional quality and data efficiency of our dataset: one does not need to pretrain a massive multimodal model from scratch; simply fine-tuning a lightweight VLM on just a 9.6K subset of the 56.9K available MIB-Silver pairs yields an evaluator that highly correlates with humans. This guarantees that MIB-Silver is a highly reusable resource for future metric alignment in the community.

4 Results

All training and evaluation are conducted on a single NVIDIA A100 GPU, with 16 vCPUs (Intel(R) Xeon(R) Platinum 8462Y+) and 251 GB of system memory. The configurations of training and evaluation processes are provided in Appendix F.

4.1 EXISTING METRICS FAIL ON MIB-GOLD

We begin by testing whether existing automatic metrics can serve as reliable proxies for human judgment on `MIB-Gold`. This is a meaningful stress test rather than an artifact of noisy annotation: after preference-consistency filtering, `MIB-Gold` retains 94.1% of seen-generator pairs and 90.4% of unseen-generator pairs, providing a stable yet challenging benchmark for metric evaluation. To establish a comprehensive performance ceiling, we therefore select a representative suite of baselines spanning four distinct paradigms: 1) **Low-level Reconstruction Metrics**: PSNR Hore & Ziou (2010), SSIM Wang et al. (2004), and MS-SSIM Wang et al. (2003) are included to measure pixel-level fidelity; 2) **Semantic Alignment Metrics**: DINOv2 Oquab et al. (2023), CLIP Radford et al. (2021), and SigLIP Zhai et al. (2023) (both Text-based and Image-based) are used to evaluate high-level semantic consistency and identity similarity; 3) **General Human Preference Models**: PickScore Kirstain et al. (2023), ImageReward Xu et al. (2023), and HPS v2.1 Wu et al. (2023), which are specifically trained on human feedback for text-to-image alignment and aesthetics; 4) **Identity-Specific Metrics**: SCR (Subject Consistency Rate) Chen et al. (2026), which focus on identity-preserving features critical for personalization.

To assess whether existing metrics can reliably evaluate multi-subject personalized generation, we conduct a systematic comparison against human pairwise preference on MIB-Gold. Specifically, we evaluate each metric on 4,020 valid pairwise comparisons, measuring pairwise agreement with double-blind human annotations, where 0.5 corresponds to random chance in a balanced binary setting. As shown in Figure 1, the results reveal a consistent failure across existing paradigms. General-purpose preference models collapse near or below random chance, PickScore achieves 0.4857, HPS v2.1 reaches 0.5201, and PSNR yields 0.3987, indicating active misalignment with human judgment. Identity-specific metrics including SCR, DINOv2, and SigLIP-I demonstrate non-trivial alignment, yet their gains remain confined to single-dimensional facets such as identity preservation or global semantic consistency. Critically, no existing metric provides a unified reasoning signal across Existence, Appearance, and Interaction simultaneously—the multi-dimensional structure underlying human preference in this setting. This confirms a fundamental evaluation gap and motivates the design of MIE.

Besides, we evaluate MIE against leading identity-assignment evaluators. As detailed in Appendix G, MIE consistently outperforms existing baselines across all MultiHuman-Testbench metrics, demonstrating its robust capacity to preserve ranking separability in multi-subject scenarios where traditional metrics fail.

4.2 MIE ALIGNS BETTER WITH HUMAN PREFERENCE

Given that existing metrics are misaligned with human preference on MIB-Gold, we next ask whether MIB-Silver can be used to train a better evaluator. Our results show that it can. Across all six exported checkpoints, the strongest variant is `qwen35_4b_lora_layer`, which achieves an overall pairwise accuracy of 0.922, including 0.982 on the seen-generator subset (Nano Banana/MOSAIC) and 0.884 on the unseen-generator subset (GPT-Image-1.5, Flux2, Seedream 4.5, and GLM). This substantially exceeds the strongest third-party baseline, showing that supervision derived from `MIB` can be translated into a materially stronger human-aligned metric.

The gains are not limited to pairwise ranking. The same 4B LoRA-layer model reaches a macro-F1 of 0.818, indicating that the evaluator is not merely learning a shallow preference score, but capturing a meaningful portion of the fine-grained diagnostic structure underlying human judgments. More importantly, this result shows not only that a larger model can rank better, but that `MIB-Silver` provides sufficiently structured supervision to train a human-aligned and diagnostically meaningful evaluator. As Figure 4 shows, the strongest model is also strongest on category-level F1, linking pairwise alignment to meaningful per-dimension supervision. This matters because the binding problem in multi-subject personalization is inherently multi-dimensional: a useful evaluator must jointly reason about existence, appearance, and interaction rather than collapse everything into a single weak aesthetic signal.

More broadly, the six-checkpoint comparison now supports a full scaling narrative. LoRA-layer variants consistently outperform their layer-only counterparts, and the largest LoRA-based model achieves the strongest overall human alignment. As shown in Figure 4, the advantage is visible not only in pairwise accuracy but also in the category-level F1 breakdown. Thus, MIB does not

only expose the failure of existing metrics; it also enables the training of a new evaluator that tracks human preference far more faithfully while preserving meaningful per-dimension diagnostics in multi-subject settings. Table 2 reports the full six-checkpoint comparison: LoRA on the unfrozen layers improves out-of-distribution generalization and diagnostic quality at every scale, motivating our choice of LoRA-layer adaptation as the primary strategy (Table 8).

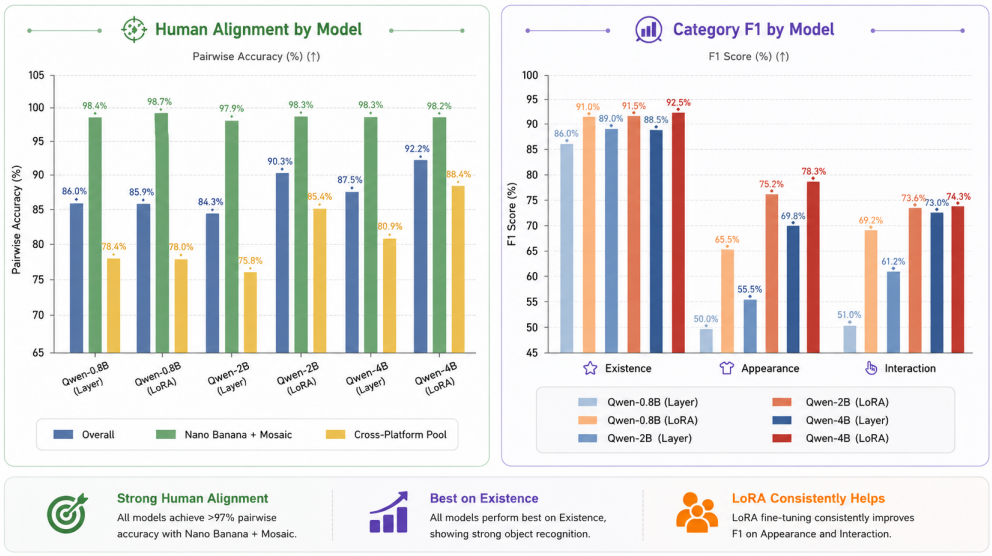


Figure 4: Human alignment of MIE variants on MIB-Gold. Left: overall, seen-generator, and unseen-generator pairwise accuracy. Right: category-level F1 across existence, appearance, and interaction. The 4B LoRA-layer evaluator is the strongest overall model and remains clearly above third-party baselines even on the unseen-generator subset.

Table 2: LoRA-layer vs. layer-only adaptation, evaluated on MIB-Gold. Δ is the gain of LoRA-layer over layer-only.

Backbone	Overall Pairwise Acc. (Δ)	Unseen Split Acc. (Δ)	Diagnostic Macro-F1 (Δ)
0.8B	-0.001	-0.004	$+0.128$
2B	$+0.061$	$+0.096$	$+0.116$
4B	$+0.046$	$+0.076$	$+0.044$

Deep Analysis and Robustness. To further validate MIE, we conducted extensive robustness checks and fine-grained analyses (detailed in Appendix H). First, zero-shot evaluation on 22 entirely novel real-world photographic subjects (10 human identities, 12 objects) confirms that MIE's alignment with human preference remains highly robust across reference domain shifts. Second, a per-subject difficulty analysis reveals that long-tailed novel categories are not the primary generative bottleneck; rather, preserving human identity under relational load remains the core challenge. Finally, a breakdown analysis across model scales confirms that our chosen LoRA-layer adaptation strategy uniquely preserves diagnostic sensitivity for Appearance and Interaction under generator shift, validating its robustness over naive parameter scaling.

Benchmarking Generators on MIB. To establish baseline performance and identify room for growth, we evaluated generator baselines using MIE on 7,806 valid images from the Gold Set. MIE evaluates each candidate independently, giving a probability in [0,1] for each of the three diagnostic dimensions (Existence, Appearance, Interaction); the columns 3-5 in Table 3 represent these probabilities averaged over the Gold Set. Because an image is only considered correctly generated when existence, appearance, and interaction are simultaneously satisfied, we also define a strict All Satisfied Rate measuring the proportion of images with all three dimensions > 0.5. As shown in Table 3, current SOTA models fundamentally struggle with multi-subject generation, with the strongest model achieving only a 52.4% success rate.

Table 3: MIE-based benchmarking of six generator baselines on the 7,806 valid Gold Set images. Existence, Appearance, and Interaction report the mean MIE diagnostic probability ($\in [0, 1]$) per dimension; All Satisfied is the fraction of images with all three dimensions > 0.5.

Generator	# Images (n)	Existence	Appearance	Interaction	All Satisfied
Seedream 4.5	1,226	**0.925**	0.674	0.477	47.2%
GPT-Image-1.5	1,226	0.846	**0.684**	0.491	**52.4%**
Nano Banana	1,451	0.811	0.549	**0.558**	42.0%
Flux.2-klein-9B	1,226	0.772	0.402	0.310	22.3%
GLM	1,226	0.402	0.148	0.079	1.6%
MOSAIC	1,451	0.281	0.121	0.065	0.8%

Existence ranks highest (reaching up to 0.925) but degrades as relational load grows; outright omission is the primary failure mode, GLM dropping up to half the subjects in eight-entity scenes. Appearance sits in the middle but collapses to 0.121 for model like MOSAIC. This occurs because existence failures are rarely isolated: models fail to maintain physical boundaries and frequently compensate for omitted subjects by fusing their specific attributes onto surviving figures, causing severe cross-subject feature bleeding and corrupted hybrid identities. Interaction is the severest bottleneck, failing to exceed 0.558 for any generator. Current architectures consistently collapse under multi-action overload. When assigned concurrent relational roles, models default to characteristic deformation errors: they either perform figure splitting by unnaturally duplicating a subject to satisfy actions independently, or they exhibit action dropout by silently discarding assigned constraints.

Limitations and Broader Impacts. While MIB provides a rigorous evaluation testbed, its current iteration relies on a controlled studio setting and a finite set of 80 reference subjects, which may not fully capture in-the-wild visual variations or long-tail cultural concepts. Methodologically, our VLM-derived silver labels contain inherent model noise, and the single-reference protocol only partially disentangles genuine identity preservation from trivial replication. Societally, although MIB sets a safer standard by avoiding scraped private data, improved personalization metrics carry dual-use risks for deepfake generation and require continuous auditing for representational biases. A comprehensive discussion of these limitations and societal impacts is provided in Appendix I.

5 Conclusion

We introduced **MIBE**, a benchmark-and-metric framework for multi-subject personalized image generation, centered on the concept of *binding*: the correct assignment of visual identities, appearances, and interactions to their designated subjects. The primary contribution of this work is **MIB**, a large-scale, rigorously annotated benchmark constructed via hierarchical prompt design and factorized scene buckets. MIB provides (i) a 60K-pair silver training corpus with VLM-derived preference labels and attribute annotations, reaching 95.1% cross-VLM preference agreement, enabling scalable metric learning and model alignment; and (ii) a 4K-pair gold evaluation set with double-blind human judgments across six state-of-the-art generators, providing a reproducible, human-grounded testbed for future comparisons. To our knowledge, MIB is the first dataset to diagnose binding failures across factorized relation types and subject counts up to eight. Building on MIB, we further propose **MIE**, a reference-conditioned evaluator that operationalizes binding quality into three interpretable dimensions, Existence, Appearance, and Interaction, alongside a scalar preference signal. MIE validates the utility of MIB as a training resource: trained exclusively on the silver set, MIE achieves 0.922 overall pairwise accuracy against human preference on the gold set, including 0.982 on seen generators and 0.884 on unseen generators, without any target-specific fine-tuning, demonstrating that the annotation protocol captures transferable binding signals rather than generator-specific artifacts. We position MIBE as open infrastructure for the multi-subject personalization community, and hope it becomes a standard testbed that accelerates progress on the binding problem and raises the bar for rigorous evaluation in personalized image generation. All data, annotations, and evaluation code will be publicly released under permissive licenses to maximize reproducibility and community uptake.

AI use statement

Generative AI is part of the research methodology of this work, and we disclose these uses as follows. (i) *Synthetic data generation*: all 80 benchmark reference images were generated with

GPT-Image-1, and all candidate images were produced by the personalized image generators under study (Nano Banana, MOSAIC, GPT-Image-1.5, Flux2, Seedream 4.5, and GLM). (ii) *Dataset labeling and cleaning*: MIB-Silver preference and diagnostic labels were produced by two VLM judges (gemini-2.5-flash and gemini-3.1-flash-lite-preview) following the SOP in Appendix D, and a lightweight LLM validator was used to check prompt constraints during prompt construction (Section 3.1.1). All MIB-Gold labels used for evaluation were produced by human annotators (Section 3.1.3). (iii) *Evaluator*: MIE is obtained by fine-tuning open vision-language backbones (Qwen3.5) on MIB-Silver (Section 3.2; Appendix F). (iv) *Code and writing assistance*: we used AI coding assistants to help write and edit experiment and figure-generation code, and AI assistants to edit the manuscript for readability and to help format the bibliography. All AI-assisted code was reviewed and run by the authors, and every bibliography entry was checked by the authors against its original source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

MIB avoids web-scraped personal photographs: all benchmark reference identities are synthetic, and the real photographs used in the reference-extension study (Appendix H) are licensed. Human annotators judged only generated images of synthetic or licensed subjects. Demographic attributes of reference subjects are reported as depicted and were annotated by the authors (Table 7). Because a stronger evaluator could also be used to optimize personalized generators, including for misuse such as multi-person deepfakes, we release MIE as a diagnostic research tool and discuss dual-use and representational-bias risks in Appendix I.

Reproducibility statement

The benchmark construction procedure, including prompt factorization and validation rules, is described in Section 3.1, and the reference pool in Appendix C. The full VLM annotation SOP is given in Appendix D. MIE's input template, read-out procedure, and dual-head design are specified in Section 3.2, and all training, loss, inference, and baseline configurations are listed in Appendix F. The evaluation protocol for the MultiHuman-Testbench baselines is detailed in Appendix G. All data, annotations, and evaluation code will be publicly released.

References

Shubhankar Borse, Seokeon Choi, Sunghyun Park, Jeongho Kim, Shreya Kadambi, Risheek Garrepalli, Sungrack Yun, Durga Malladi, and Fatih Porikli. Multihuman-testbench: Benchmarking image generation for multiple humans. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, 2025. doi: 10.52202/085713-3491.

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. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 7437–7459, 2025. doi: 10.52202/085713-0254.

Zhihan Chen, Yuhuan Zhao, Yijie Zhu, and Xinyu Yao. When identities collapse: A stress-test benchmark for multi-subject personalization. *arXiv preprint arXiv:2603.26078*, 2026.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Divake Kumar, Sina Tayebati, Devashri Naik, Ranganath Krishnan, and Amit Ranjan Trivedi. Vlm judges can rank but cannot score: Task-dependent uncertainty in multimodal evaluation. *arXiv preprint arXiv:2604.25235*, 2026.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941. IEEE, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8640–8650. IEEE, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Yuta Oshima, Daiki Miyake, Kohsei Matsutani, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Multibanana: A challenging benchmark for multi-reference text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 448–460, 2026.

Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. Dreambench++: A human-aligned benchmark for personalized image generation. In *International Conference on Learning Representations*, volume 2025, pp. 46010–46032, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510. IEEE, 2023.

Wenqing Tian, Hanyi Mao, Zhaocheng Liu, Lihua Zhang, Qiang Liu, Jian Wu, and Liang Wang. Multibind: A benchmark for attribute misbinding in multi-subject generation. *arXiv preprint arXiv:2603.21937*, 2026.

Shulei Wang, Longhui Wei, Xin He, Jianbo Ouyang, Hui Lu, Zhou Zhao, and Qi Tian. Psr: Scaling multi-subject personalized image generation with pairwise subject-consistency rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14756–14766, 2026.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. Withanyone: Toward controllable and id consistent image generation. In *International Conference on Learning Representations*, volume 2026, pp. 130581–130610, 2026.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Zhu Xu, Zhaowen Wang, Yuxin Peng, and Yang Liu. Interact-custom: Customized human object interaction image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9500–9508, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

A Appendix: Human Annotation Observations and Failure Mode Analysis

During the human annotation process for the gold test set, annotators systematically recorded recurring failure patterns across evaluated image pairs. These observations surfaced three consistent and practically significant findings, which we document here to inform future dataset construction and model development efforts.

Table 4 reports the full breakdown across ethnicity, age band, and wardrobe (gray T-shirt, jeans), for MIB-Silver and MIB-Gold overall and by subject count N in MIB-Silver; co-occurrence of similar subjects roughly doubles from $N = 2$ to $N = 8$ across every axis. This design directly targets our Appearance dimension: when two visually or demographically similar subjects share a scene, generators that rely on coarse cues such as ethnicity, hairstyle, or clothing rather than fine-grained facial identity are exposed, and an identity mix-up is scored as an Appearance failure (wrong face, hair, or garment attributed to the wrong subject, including cross-subject feature bleeding; Appendix A), while Existence separately catches the more severe case where two similar subjects collapse into one.

Table 4: Co-occurrence rate of visually or demographically similar human subjects within a scene, overall in MIB-Silver and MIB-Gold, and by subject count N in MIB-Silver.

Metric	Silver	Gold	$N = 2$	$N = 4$	$N = 6$	$N = 8$
Same ethnicity	39.4%	38.9%	7.3%	26.4%	54.7%	69.0%
Same age band	47.4%	47.1%	10.8%	35.3%	67.4%	76.3%
Both gray T-shirt	16.9%	16.9%	2.6%	9.5%	21.0%	34.4%
Both jeans	41.7%	41.3%	10.4%	30.3%	57.2%	69.0%

All references are standardized as isolated single-subject studio images at 512×512, allowing the same identity anchors to be recombined across prompts. For each prompt-reference task, a personalized generator receives the guiding prompt and reference images as visual conditioning inputs, producing a candidate image. MIB forms pairwise comparisons by placing two candidates under the same prompt-reference condition, so labels capture relative generation quality rather than prompt difficulty. References are generated separately from all candidate generators to reduce generator-specific leakage.

Failure Rates Increase with Subject Count. Annotators consistently found that images depicting a larger number of subjects were significantly harder to evaluate correctly, and that generative models performed more poorly on such cases. The most common failure modes in high-subject-count scenes were subject omission, where one or more individuals were missing from the generated output, and appearance drift, where the physical attributes of individual subjects degraded as the total number of subjects increased. This trend was particularly pronounced for visually similar subjects (e.g., individuals wearing comparable outfits or sharing similar physical features), where models frequently confused or merged identities. Facial attributes and clothing details were identified as the two most error-prone visual channels under these conditions.

Existence Failures Frequently Co-occur with Appearance Failures. A substantial proportion of annotated samples exhibited a systematic co-occurrence between Existence and Appearance failures. When a model fails to render all requested subjects as distinct entities, it often compensates by blending the features of missing subjects into the remaining ones, for instance, combining the facial identity of subject A with the wardrobe of subject B into a single generated figure. This entanglement means that Existence failures are rarely isolated: they tend to cascade into cross-subject feature bleeding, inflating Appearance failure rates in a correlated rather than independent manner. This observation motivates treating the three diagnostic dimensions as structured but not fully orthogonal signals in practice, despite their conceptual independence.

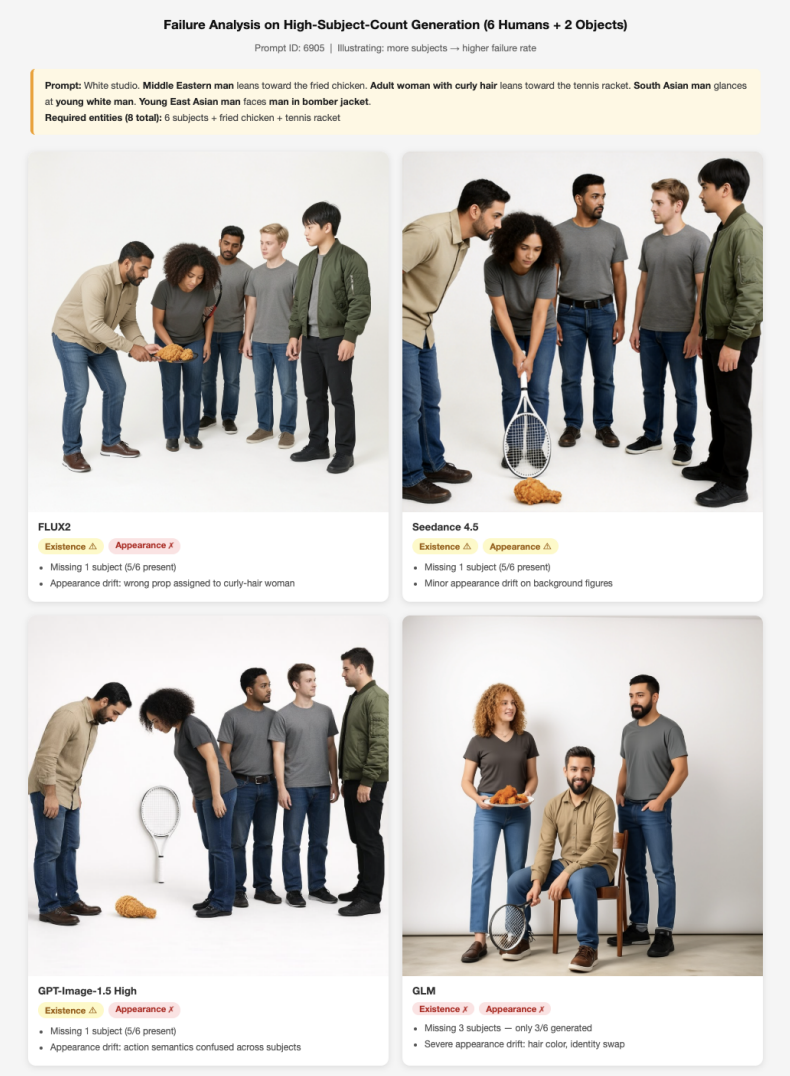


Figure 5: **Failure Rates Increase with Subject Count** A representative 6-subject, 2-object prompt (ID: 6905) evaluated across four generative models. All models exhibit Existence failures, with subject omission ranging from 1 missing subject (FLUX2, Seedream-4.5, GPT) to 3 missing subjects (GLM).

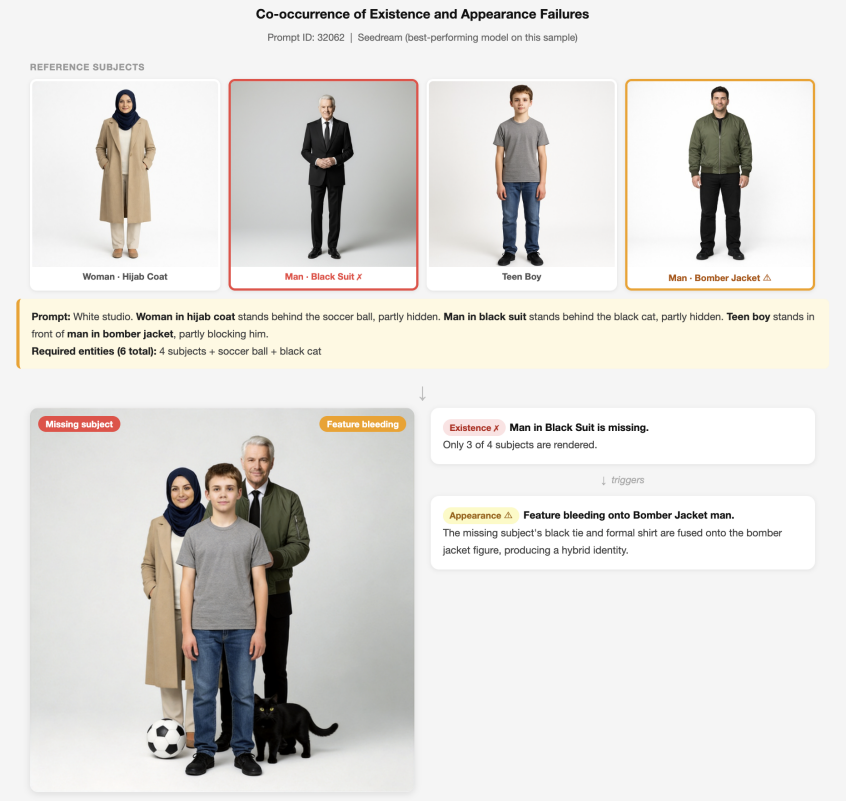


Figure 6: **Existence Failures Frequently Co-occur with Appearance Failures** A representative 4-subject prompt (ID: 32062) generated by Seedream, the best-performing model on this sample. The missing subject (Man in Black Suit) is not simply absent: its visual attributes, the black tie and formal shirt, are redistributed onto the surviving Bomber Jacket figure, producing a hybrid identity.

Multi-Action Overload Leads to Subject Deformation Annotators consistently observed that when a single subject is assigned more than one concurrent action or interaction, generative models struggle to faithfully execute all specified behaviors. This overload manifests in two characteristic failure patterns. First, *figure splitting*: the model resolves the conflict by duplicating or fragmenting the subject, distributing distinct actions across two visually similar but non-identical figures, effectively cloning the identity to satisfy each action independently. Second, *action dropout*: the model selectively ignores or partially executes one of the assigned actions, producing a subject that satisfies only a subset of the prompt specification. Both failure modes were reproducible across multiple models and prompt configurations, suggesting that multi-action binding exceeds the compositional capacity of current generative architectures rather than reflecting prompt ambiguity.

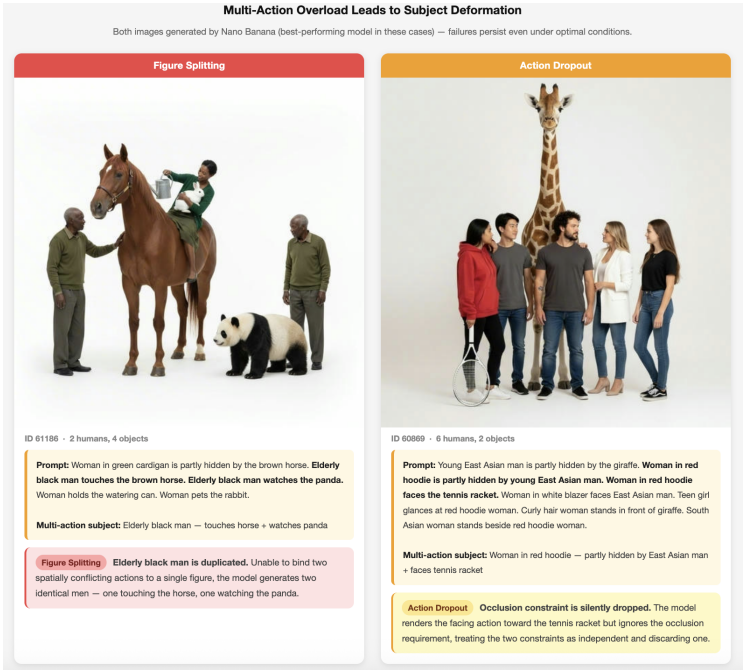


Figure 7: **Multi-Action Overload Leads to Subject Deformation.** When a subject is assigned more than one concurrent action, models either clone the subject to satisfy each action independently (Figure Splitting, ID: 61186) or silently discard one of the assigned actions (Action Dropout, ID: 60869). Both failure modes are observed in Nano Banana, a better performing model in these cases.

A.1 Structured Example Grid

To make the compositional design of MIB concrete, Table 5 presents representative prompts sampled from a 3×2 slice of our full factorization grid: three relation categories (no interaction/no occlusion, occlusion without interaction, and occlusion with physical interaction) crossed with two subject counts (2 and 8). Each row reuses the same shared subject vocabulary and spatial-relationship phrasing, with only the entity count and relation density varying across columns.

This structured format clarifies that our prompts are compositional, recombining a shared subject vocabulary under a controlled grid where spatial relationships (e.g., occlusion) are explicitly stated. This visually confirms that MIB's difficulty stems from the density of simultaneous relational constraints rather than from uncontrolled prompt variation.

B Appendix: Summarized Table of Related Work

Existing benchmarks for personalized and multi-subject image generation have made significant progress, yet they collectively leave a critical evaluation gap that MIB is designed to address. As summarized in Table 6, earlier text-to-image benchmarks such as T2I-CompBench Huang et al. (2023) entirely lack reference image support, making them unable to measure subject identity fidelity. Single-subject benchmarks like DreamBench Ruiz et al. (2023) and DreamBench++ Peng et al. (2025) introduced reference-based evaluation but restrict testing to one concept at a time, leaving multi-subject binding behavior unexamined. More recent efforts including XVerseBench Chen et al. (2025), PSRBench Wang et al. (2026), OmniContext Wu et al. (2025), MultiBanana Oshima et al. (2026), and MultiHuman-Testbench Borse et al. (2025) extend evaluation to multiple subjects, yet none provides factorized relation-level stress testing that isolates failure modes arising from physical interaction, occlusion, and identity leakage under increasing subject count. MultiBind Tian et al. (2026) targets attribute-level misbinding diagnostics but does not consider the geometric and occlusion-driven binding failures that emerge in contact-rich scenes. WithAnyone Xu et al. (2026) and its MultiID-Bench evaluate multi-person identity consistency and natural pose or expression

Interaction Class	2 Subjects	8 Subjects
No Interaction & No Occlusion	*"White studio. elderly east asian woman stands near the headphones."*	*"White studio. elderly east asian woman stands near the headphones. teen girl looks at the donut. man black suit stands beside the toolbox. south asian woman leans toward the travel suitcase."*
Occlusion without Interaction	*"White studio. south asian woman is partly hidden by young white woman."*	*"White studio. young white woman is partly hidden behind the zebra. south asian woman is partly hidden by young white woman. south asian woman stands near the succulent plant. young east asian woman glances at south asian woman."*
Occlusion & Physical Interaction	*"White studio. The tennis racket is partly hidden by the zebra."*	*"White studio. elderly black man is partly hidden by the zebra. middleaged black man feeds the zebra. middleaged black man watches the duck. elderly black man twirls the tennis racket."*

Table 5: Representative prompts from a 3×2 slice of MIB's factorized grid (relation type $\times$ subject count). The shared subject vocabulary and explicit spatial-relationship phrasing (e.g., "partly hidden by") are held constant while entity count and relation density scale up, isolating relational load as the source of increasing task difficulty.

variation via face/identity metrics, but remain centered on human subjects and do not provide the decoupled Existence/Appearance/Interaction diagnostics or non-human, occlusion, and physical-interaction stress that MIB targets. Crucially, no prior benchmark provides human pairwise gold evaluation labels, which are essential for validating automated metric reliability in the multi-subject setting. MIB addresses all of these gaps simultaneously: it introduces factorized relation-level binding stress tests across up to eight subjects, pairs scalable silver supervision with human gold pairwise preference labels, and provides the first diagnostic framework specifically designed to expose and quantify identity collapse and appearance leakage in complex multi-subject generation. Built on this factorized dataset, MIE further pairs a comprehensive human-aligned global score with these same decoupled diagnostics: the score is trained to align with human pairwise preference and is evaluated against MIB-Gold's 4,020 double-blind human comparisons, which contribute no training signal, so MIBE is the only benchmark–evaluator pair to offer both a validated human-aligned ranking and interpretable per-dimension diagnostics that prior datasets and metrics cannot reliably extract.

Benchmark	Year	Ref. Images	Multi-subj.	Stress / Eval. Design	Human Gold Eval.	Pref./Diag. Supervision	Main Gap
T2I-CompBench Huang et al. (2023)	2023	×	Text-only	Text-only composition	×	×	Cannot measure reference identity, identity collapse, or appearance leakage.
DreamBench Ruiz et al. (2023)	2023	✓	×	Single-subject prompts	×	×	No multi-subject binding or interaction evaluation.
DreamBench++ Peng et al. (2025)	2024	✓	×	Single-subject, human-aligned automated eval.	×	×	Broader single-subject evaluation, but no high-subject-count binding stress.
CustomConcept101 Kumari et al. (2023)	2023	✓	Limited	Multi-concept composition	×	×	Not designed for contact-rich or occlusion-driven identity binding.
XVerseBench Chen et al. (2025)	2025	✓	✓	Single / dual / triple subjects	×	×	Limited high-subject-count and relation-level stress factorization.
PSRBench Wang et al. (2026)	2025	✓	✓	Broad capability subsets	×	Reward metrics	Capability-oriented; not a factorized diagnostic benchmark for interaction/occlusion binding failures.
OmniContext Wu et al. (2025)	2025	✓	✓	In-context character / object / scene eval.	×	Automated judge scores	Broad in-context consistency benchmark, but lacks human gold and factorized relation-level binding stress.
MultiBanana Oshima et al. (2026)	2025	✓	✓	Multi-reference challenges	×	VLM-based scores	Covers multi-reference difficulty, but not hierarchical interaction/occlusion diagnosis or human pairwise gold labels.
MultiHuman-Testbench Borse et al. (2025)	2025	✓	✓	Multi-human actions / faces	×	Metric suite	Restricted to human subjects; does not cover human-object or object-object binding.
MultiBind Tian et al. (2026)	2026	✓	✓	Attribute / slot misbinding	×	Diagnostic metrics	Focuses on attribute-level misbinding rather than factorized physical interaction and occlusion stress.
WithAnyone Xu et al. (2026)	2025	✓	✓	Human ID-consistency & pose/expression variation	×	Face/identity metrics	Centered on human/character identity fidelity; no non-human subjects, occlusion/relation stress, or decoupled diagnostics.
MIB (Ours)	**2026**	✓	✓	**Factorized relation-level binding stress**	✓	✓	–

Table 6: Comparison with existing benchmarks. DreamBench and CustomConcept101 refer to evaluation sets from DreamBooth and Custom Diffusion respectively. MIB uniquely targets factorized relation-level binding stress with both scalable silver supervision and human gold evaluation.

C Appendix: MIB Reference Subjects

MIB uses a fixed pool of 80 GPT-Image-1 reference subjects, including 30 human and 50 non-human identities. Human references vary in age, gender presentation, ethnicity, clothing, hairstyle, and appearance, explicitly spanning a wide range of ages (teens to elderly), diverse ethnicities, unique personal attributes (e.g., curly, braided, or bald hairstyles), and specific cultural attire (e.g., hijabs or distinctive coats); to introduce deliberate similarity, half of the subjects share near-identical casual wear (gray T-shirts and jeans), so identity cannot simply be read off clothing, and visually similar subjects further co-occur within prompts, forcing models to preserve fine-grained facial features rather than relying on attire shortcuts. The 50 non-human references cover long-tailed classes spanning animals, foods, tools, and objects, including occluders. Table 7 summarizes the demographic and category composition of the 80-subject reference pool.

Seven references (4 large animals, 3 large objects) additionally serve as valid person-occluders in occlusion prompts, and the wardrobe overlap described above is realized concretely: 39.4% of Silver scenes contain a same-ethnicity human pair, rising to 69.0% at eight subjects. Two Middle Eastern subjects specifically (e.g., a *middle eastern man* paired with a *middle eastern woman in hijab*) co-occur in 2.2% of Silver scenes, rising to 4.9% at eight subjects, directly instantiating this same-ethnicity case.

We use synthetic, GPT-Image-1 generated references rather than real photographs for two reasons. First, releasing real human photographs at benchmark scale raises copyright and consent concerns that would compromise full reproducibility and public redistribution; synthetic references avoid this while following recent personalization benchmarks for fair comparability, and all synthetic images are kept strictly photorealistic so that identity preservation is not trivialized into an easy proxy task. Second, several stress conditions in MIB, such as same-colored clothing or matching hairstyles, require subjects that differ only along one targeted axis, combinations that are difficult to source from real photographs at scale but straightforward to construct precisely with generated references. Assembling an equivalent pool from licensed real photographs would also force acceptance of that corpus's own demographic and category imbalances, confounding the very axes MIB is designed to isolate. Generated references further provide a single canonical view per identity that stays fixed across every task instance in which that identity appears, giving Appearance fidelity a stable, neutral anchor.

Table 7: Demographic and category distribution of the 80-subject reference pool (30 human, 50 non-human). All human identities are synthetic; age and ethnicity are as depicted, annotated by the authors.

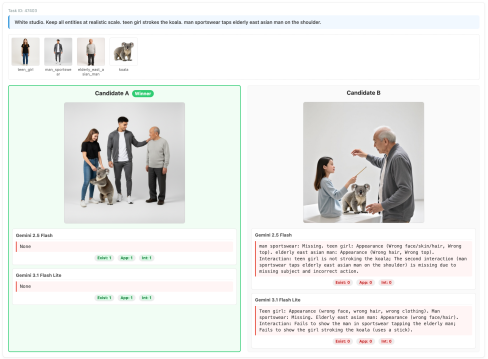
Humans (30)	
Gender	15 male / 15 female
Age band	2 teens, 6 young adults (18–30), 15 adults (30–60), 7 elderly (60+)
Ethnicity (as depicted)	11 White, 7 Black, 6 East Asian, 3 South Asian, 3 Middle Eastern
Hair style	15 short, 5 long straight, 3 bob/shoulder-length, 3 curly/coiled, 1 braided, 1 bald, 2 hijab-covered
Hair color	∼15 black/dark, ∼6 brown, 2 light/blond, 7 gray/white
Facial features	3 with facial hair, 2 hijab, 1 bald
Tops	9 gray T-shirts, 3 gray sweaters, 18 distinctive tops (flannel, hoodie, blazer, suit, dress, coat, cardigan, colored tees)
Bottoms	17 jeans, 6 dark trousers, 3 khaki chinos, 4 other (joggers/skirt/dress/cream trousers)
Non-humans (50)	
Animals (20)	10 wild mammals, 6 domestic/farm, 4 birds
Foods (10)	7 savory, 2 desserts, 1 fruit
Objects (20)	5 sports equipment, 4 electronics/photography, 4 household, 6 travel/tools, 1 musical instrument

D APPENDIX: SOP PROMPT

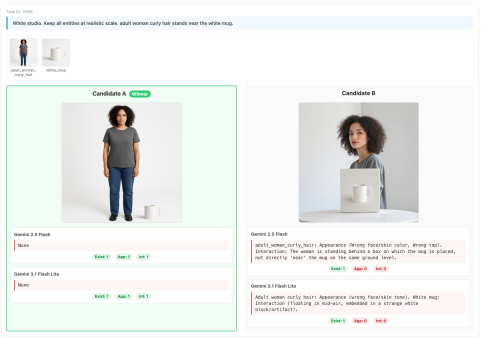
```
You are an expert judge for multi-subject personalized image generation.

You will receive:
1. The original generation prompt.
2. Reference images for each subject mentioned in the prompt.
3. Candidate image A.
4. Candidate image B.

Your job consists of two parts:
Part 1: Independent Evaluation
Evaluate candidate A and candidate B independently. For each image,
strictly evaluate:
- Existence: Are ALL the required subjects in the references physically
present?
  (Generic versions count as 1; completely missing counts as 0).
- Appearance: Are the subjects rendered WITHOUT structural deformations?
  Do the generated subjects strictly match the references across
  ALL features: Face/Skin tone, Hair style, Top clothing, AND Bottoms (pants)?
- Interaction: Does the generated image reflect the described actions, maintain
  correct proportions, and follow realistic spatial/physical laws?

Part 2: Overall Comparison
Compare Candidate A and B and decide the overall winner.

OUTPUT RULES:
- Use ONLY the provided prompt and images.
- All dimensional scores ('*_existence', '*_appearance', '*_interaction') must be
  strictly binary: 1 (fully correct) or 0 (any issue exists).
- 'winner' must be either "A" or "B" (Never "Tie").
- Return valid JSON ONLY. Absolutely NO markdown formatting.
- CRITICAL: To focus on errors, the JSON MUST begin with `a_flaw_log` and
  `b_flaw_log`. DO NOT list perfect subjects. ONLY list subjects with errors and
  explicitly state what is wrong (Missing, Wrong Face/Hair/Skin, Wrong Top,
  Wrong Pants, Action/Spatial/Physics Error). If an image is 100% perfect,
```

```
  write "None".

SCORING NOTES & SPECIFIC EDGE CASES:
- Appearance (Strict Identity & Wardrobe): You MUST check the whole person. If a
  subject has the correct top but wrong pants, wrong hair, or mismatched skin
  tone/face compared to the reference, score Appearance as 0 and log the
  specific flaw (e.g., "Woman red hoodie: Wrong face/skin color",
  "Man flannel shirt: Wrong pants color").
- Appearance (Mangled/Melted): Any structural artifacts (extra limbs, mangled
  faces) = Appearance 0.
- Existence vs Appearance: If requested "man denim jacket" is drawn as a generic
  man in a grey sweater, Existence is 1, Appearance is 0.
- Interaction (Partial): Evaluate interactions based ONLY on rendered subjects.
  If an interaction requires a missing subject, Interaction is 0.
- Interaction (Physics, Environment & Scale): Obvious physical violations
  (e.g., objects floating in mid-air without support) OR severe scale distortions
  between objects (e.g., abnormally giant bread, miniature people) = Interaction 0.
  HOWEVER, the presence of unprompted but logical supporting structures
  (e.g., tables, pedestals) used to hold objects is PERMITTED and should NOT be
  penalized.

EXPECTED JSON SCHEMA:
{
  "a_flaw_log": "Middle eastern man: Appearance (wears green sweater instead of
  beige shirt). Man flannel shirt: Appearance (wrong pants color). Woman red
  hoodie: Appearance (mismatched face/skin/hair). Interaction: Man is not
  occluded by the woman; The burger is floating in mid-air.",
  "b_flaw_log": "Middle eastern man: Missing. Man bomber jacket: Missing. DSLR
  camera: Missing. Interaction: Man is not walking towards/staring at the woman;
  Fails due to missing interaction targets.",
  "a_existence": 1,
  "a_appearance": 0,
  "a_interaction": 0,
  "b_existence": 0,
  "b_appearance": 0,
  "b_interaction": 0,
  "winner": "A"
}
```

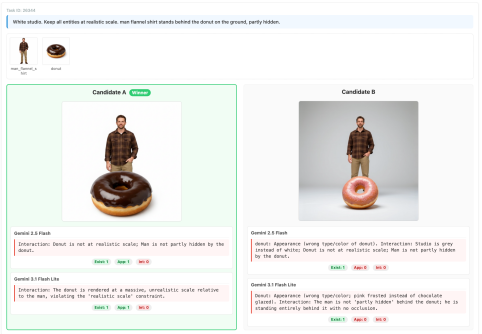
E APPENDIX: DEMOS OF GENERATED TRAINING LABELS



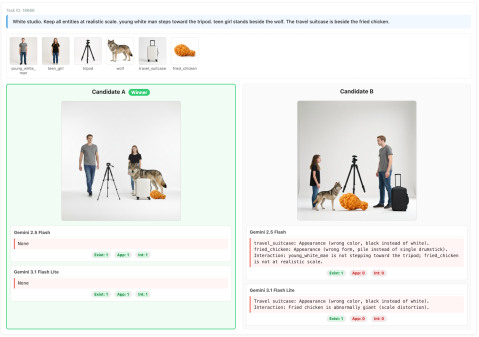
(a) Action and Target Errors



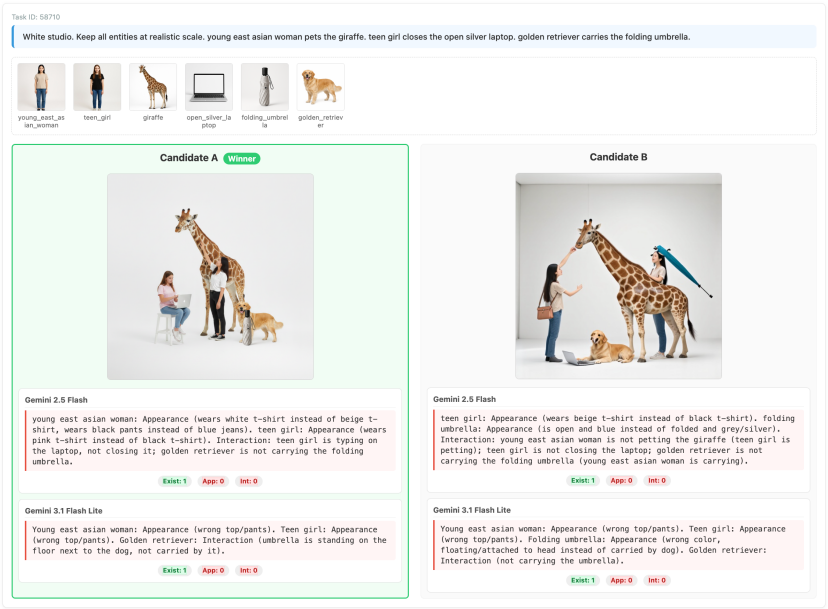
(b) Physics and Spatial Violations



(c) Appearance and Occlusion Errors



(d) Scale Distortion and Color Mismatch



(e) Robust Diagnostics in Highly Complex (6-Subject) Scenes

Figure 8: Qualitative examples of strict cross-model consensus. The externalized flaw logs demonstrate that both VLMs effectively utilize the SOP to catch fine-grained errors in appearance, physical realism, interaction semantics, and complex multi-subject scaling.

F MIE CONFIGURATION

Table 8: Training Hyperparameters (shared by all six MIE variants, with layer-only variants omitting LoRA; the 0.8B variant is shown).

Parameter	**Value**
Base Model	Qwen3.5-0.8B (via unsloth)
Adaptation Method	LoRA ($r = 16$, $\alpha = 32$, dropout $= 0$, no bias) on attention/MLP
Additional Unfreezing	Top 4 backbone layers
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.01
LR Scheduler	Cosine schedule with warmup
Warmup Ratio	0.03
Batch Size (per device)	1
Gradient Accumulation Steps	16
Target Effective Batch Size	16
Training Budget	600 optimizer steps ($\approx$9.6K pairs)
Gradient Clipping Norm	1.0
Image Resolution	512×512
Global Random Seed	3407

Table 9: Multi-task Loss Configuration

Component	**Setting**
Preference Ranking Loss	Margin Ranking Loss (margin $= 0.1$)
Ranking Loss Weight (α)	1.0
Diagnostic Classification Loss	BCEWithLogitsLoss
Classification Loss Weight (β)	0.5
Early Stopping	Patience: 1 epoch, min_delta $= 10^{-3}$

Table 10: Inference and Evaluation Setup

Setting	**Detail**
MIE Inference Precision	Default checkpoint precision (bf16/fp16)
Evaluation Resolution	512×512 (preserving aspect ratio)
Metric 1: Preference Accuracy	Binary choice based on scoring head logits
Metric 2: Diagnostic Output	Weakest dimension via Sigmoid from classification head

Table 11: Baseline Models Configuration

Model	Configuration
CLIP Baseline	`openai/clip-vit-base-patch32` Text Truncation: Enabled Scoring: Cosine similarity between text embedding and pooled image embedding
DINO Baseline	`facebook/dinov2-base` Scoring: Cosine similarity between the average normalized subject reference embeddings and generated image embeddings

G Appendix: Detailed Comparison with MultiHuman-Testbench

Here, we additionally compare MIE against the three metrics proposed by MultiHuman-Testbench Borse et al. (2025): Hungarian identity similarity, which assigns generated faces to reference identities via optimal bipartite matching and penalizes unmatched or duplicated identities; action score, which verifies whether a described action is depicted; and count accuracy, which compares the number of detected human faces against the target subject count.

Following the same pairwise protocol as Figure 1, a metric agrees on a pair when it scores the human-preferred candidate higher, with ties and non-computable scores counting as half agreement. Because Hungarian identity similarity and action score are undefined without a detectable reference face, and count accuracy trivially rewards zero detected faces on object-only prompts, we restrict this comparison to MIB-Gold's human-consensus pairs containing at least one human subject. Reference images used in the MultiHuman-Testbench Borse et al. (2025) baseline comparison are processed with RetinaFace for face detection; if no face is detected in a reference image, the corresponding sample is skipped for both Hungarian identity similarity and action score computation. For count accuracy, the metric compares the number of detected human faces in the generated image against the target subject count, so on object-only prompts, where the target human count is zero, a generated image with zero detected faces is counted as correct. To be noted, the scores reported in Table 12 are pairwise human-preference alignment scores rather than raw metric averages, therefore, these edge cases are handled through a missing-score policy: for each human-annotated pair, an agreement between the metric-predicted winner and the majority human preference receives 1 point, a disagreement receives 0 points, and a metric tie or a missing metric score receives 0.5 points, with human ties excluded from the denominator. Under this policy, object-only prompts that yield missing MultiHuman-Testbench identity or action scores contribute 0.5 to the overall alignment score reported in Table 12.

Table 12: Pairwise agreement with majority human preference on MIB-Gold's human-subject pairs, comparing MIE against MultiHuman-Testbench's identity-assignment metrics Borse et al. (2025).

Metric	Overall	Seen	Unseen
Hungarian identity similarity	0.806	0.723	0.857
Action score	0.767	0.822	0.732
Count accuracy	0.601	0.578	0.615
MIE (ours)	0.922	0.982	0.884

MIE leads on every split, and no single MHTB metric is stable across regimes: action score is strongest on seen prompts (0.822) but drops to 0.732 on unseen prompts, while Hungarian identity similarity shows the opposite pattern (0.723 seen, 0.857 unseen). This complementary failure pattern is the clearest evidence for our core argument: a single-axis metric tracks human preference only when that axis happens to decide the pair. On the seen split, generators have comparable per-face fidelity, so raw identity similarity is close to uninformative and the human verdict instead turns on which subject is missing or whether the interaction binds correctly; the unseen split, in contrast,

contains many pairs with gross identity collapse, which an assignment-based score is well-suited to catch. The same pattern holds with density: because Level-8/6/4/2 prompts are constructed as strict nested subsets differing only in subject count, Hungarian identity similarity, computed under our original unfiltered protocol, falls from 0.806 at $N = 4$ to 0.589 at $N = 8$, approaching chance and reproducing the degradation trend of Figure 1 on a stronger identity metric; MIE alone remains above 0.88 in both regimes. Each MHTB metric also carries a structural blind spot. Hungarian identity similarity reduces face assignment to a single aggregate scalar: it is undefined whenever no human face is detected (50 of MIB's 80 reference subjects are non-human, and object-heavy or balanced prompt buckets contain few or no faces), it is invariant to relation direction (e.g., "A walks toward B" and "B walks toward A" score identically), and it cannot separate a single omitted subject from all subjects present but degraded. Action score verifies that an action is depicted, not who performs it on whom, so it is invariant to role and target inversions and to missing-target interactions, which dominate MIB's failure cases. Count accuracy is identity-agnostic by construction: an image that matches the target count via duplicated faces or a generic substitute scores perfectly, while the human-preferred image with correct identities but one omission scores as incorrect; this is starkest at high density, where count accuracy ties 91.5% of pairs on the seen split at $N = 8$, unable to rank exactly the settings MIB targets.

H Appendix: Robustness Checks and Breakdown Analysis

MIE Generalizes to Unseen Reference Classes To test whether MIE's behavior survives leaving its training inventory, we conducted a reference-extension study using structured prompts identical to those in MIB, re-rendered against a completely new source of references: licensed real-world photographs. This introduces 22 entirely novel subjects with zero overlap with MIB's original 80 reference subjects, consisting of 10 human identities with new demographic and styling combinations (e.g., a Black teen girl with braids, a Latino man in a black polo) and 12 non-human categories entirely absent from MIB (e.g., an alpaca, a desk globe, a violin, a picnic basket). MIE was not retrained for this experiment, so results strictly measure zero-shot generalization. Evaluated on the unseen-generator batch (Flux.2-klein-9B vs. GPT-Image-1.5), MIE achieves comparable AUC against human pairwise preference on both the new real-world photographs and the original synthetic references (Table 13), and its per-image Existence predictions maintain a highly robust correlation with human annotations across both reference domains, confirming that MIE's alignment with human judgment does not degrade when the reference domain shifts from synthetic studio images to real photographs.

Table 13: MIE alignment with human preference across reference domains, evaluated on the unseen-generator batch (Flux.2-klein-9B vs. GPT-Image-1.5).

Reference Source	# Images (n)	AUC	Existence Correlation
Real photographs	173	0.827	+0.932
GPT-Image-1 (synthetic)	176	0.851	+0.926

The Reference-Source Gap Is Generator-Specific and Concentrated in Appearance. Beyond aggregate AUC, we examine whether real photographic references are intrinsically harder to preserve than synthetic ones. Across the two reference sources we collected 864 pairwise comparisons (1,728 ratings from 8 annotators, 2 per pair) covering 1,722 of 1,728 candidate images (six real-reference Flux.2 renders were missing), scored with the same MIE checkpoint reported in the main paper without retraining. Real references are measurably harder, but the effect is generator-specific and concentrated in the appearance facet rather than existence: for Nano Banana, the human-marked appearance-success rate drops from 0.472 with synthetic references to 0.306 with real references, while existence success is essentially unchanged (0.896 real vs. 0.889 synthetic). Despite this generator-specific difficulty gap, MIE's evaluation validity is unaffected: per-facet existence correlation with human judgment remains essentially invariant across reference sources, and relative generator rankings are identical for both settings, reinforcing the AUC parity reported above. We note that appearance is the facet with the highest human subjectivity (Fleiss' κ = 0.19–0.34 across annotators), which sets a natural ceiling on achievable correlation and is a constraint of the annotation protocol rather than of MIE itself.

Long-Tailed Classes Are Not the Bottleneck To isolate the specific difficulty of long-tailed classes, we computed a per-subject difficulty index measuring how much a subject's presence raises or lowers overall image score relative to the baseline average (positive: easier to generate; negative: more failures), separately for the 12 novel non-human categories and 10 novel human identities introduced above (Table 14). Both human annotators and MIE independently score the novel non-human categories above the overall average, indicating that category novelty itself carries no inherent penalty; the binding difficulty instead remains concentrated on human identity preservation. MIE accurately reproduces the human difficulty index on the 22 unseen subjects (Spearman $\rho = +0.92$ for Existence), confirming that the fundamental bottleneck for current generators is not object category rarity but the binding of human facial and clothing attributes under complex relational loads.

Table 14: Per-subject difficulty index (positive: easier than baseline; negative: harder than baseline) for the 22 unseen subjects, split by category.

Metric	New Non-Human (12)	New Human Identity (10)	Contrast
Human-marked appearance	+0.0145	-0.0145	+0.029
Human-marked existence	+0.0209	-0.0208	+0.042
MIE appearance	+0.018	-0.018	+0.036
MIE existence	+0.029	-0.028	+0.057

MIE Breakdown Analysis We analyze where the gains of the learned evaluator actually come from. A consistent pattern emerges across all six checkpoints: every model performs better on the seen-generator subset than on the unseen-generator subset, confirming that cross-generator generalization remains a non-trivial challenge. However, the size of this gap depends strongly on the tuning regime. The smallest seen-to-unseen drop is achieved by `qwen35_4b_lora_layer` at -0.098, whereas the largest drop appears in `2b layer_only` at -0.221. This shows that additional model capacity alone is not enough; how that capacity is adapted is equally important.

The LoRA-versus-layer comparison reinforces this conclusion. At 2B, LoRA-layer tuning improves pairwise accuracy by 0.061 and macro-F1 by 0.116 relative to `2B layer_only`. At 4B, it still yields gains of 0.046 in pairwise accuracy and 0.044 in macro-F1 over `4B layer_only`. Even at 0.8B, where the pairwise gain is nearly neutral, LoRA-layer tuning improves macro-F1 by 0.128. These results suggest that the primary benefit of LoRA-layer tuning is not merely stronger ranking, but consistently sharper diagnostic discrimination. In other words, the full six-checkpoint comparison supports a benchmark-utility claim rather than a simple "bigger is better" story.

The category-level breakdown points to the same conclusion. `Existence` remains the easiest dimension overall, whereas `Appearance` and `Interaction` provide the more discriminative diagnostic challenge. As shown in Figure 9, generator shift is not uniform across dimensions: the strongest checkpoint remains highly stable on `Existence`, while the more semantic categories continue to differentiate model quality and adaptation strategy. Overall, the breakdown analysis clarifies that the best-performing evaluator is not simply the largest model, but the model that combines additional capacity with the right adaptation strategy and preserves diagnostic sensitivity under distribution shift.

I Appendix: Limitations and Societal Impacts

Scale and Diversity of the Gold Set. While our 4K gold evaluation set was curated under a rigorous double-blind protocol, it inherently captures a finite subspace of the vast combinations of subjects and interactions. The current benchmark relies on 80 unique reference subjects, which, despite covering major categories of humans, animals, and everyday objects, may not fully represent the long-tail distribution of real-world identities or culturally specific concepts. The white-studio setting and global realistic-scale constraint further narrow the visual distribution to controlled compositional stress, leaving in-the-wild backgrounds and stylistic variations beyond the current scope. Future iterations will expand the reference concept library along long-tail axes and extend evaluation to less constrained visual settings.



Figure 9: Breakdown analysis of MIE variants. Left: seen-to-unseen generalization gap. Middle: LoRA-layer gains over layer-only tuning at different model scales. Right: category-level F1 for the strongest checkpoint. The results show that LoRA-layer tuning improves diagnostic quality across scales, while generator shift affects diagnostic categories non-uniformly: `Existence` remains the most stable dimension, and `Appearance`/`Interaction` remain more discriminative.

Temporal Snapshot of Baseline Generators The MIB gold set reflects the performance of six representative generators at the time of collection. Given the rapid velocity of the field, newer architectures may encounter unique failure modes, such as specific multi-modal reasoning bottlenecks, not fully captured by the current distribution. To ensure the benchmark's longevity, we commit to releasing versioned extensions as transformative models emerge, providing a standardized re-annotation protocol to facilitate community-driven growth.

VLM-Induced Noise in the Silver Set The 60K silver set utilizes a VLM-based pipeline for scalable preference pseudo-labeling. Despite its utility, VLM labeling is susceptible to model-specific biases, such as the tendency to conflate surface-level image fidelity with structural binding accuracy. Furthermore, these models may underperform on culturally specific concepts underrepresented in their pretraining data. We recommend that practitioners using the silver set for model alignment (e.g., via DPO) treat these labels as high-fidelity but noisy proxies, incorporating human-in-the-loop auditing for safety-critical applications.

Single-Reference Identity Preservation vs. Replication. A single reference image per subject cannot fully distinguish genuine identity preservation across pose, expression, and lighting variation from trivial replication, the "copy-paste" failure mode in which a model directly reproduces the reference face rather than preserving identity under natural variation Xu et al. (2026). Two aspects of MIB's design partially mitigate this: Appearance is scored by human/VLM judgment rather than by feature-similarity metrics (e.g., ArcFace, DINO) that would reward exact replication, and Appearance and Interaction are rated as separate facets, so a copy-pasted subject placed in an action-requiring scene is penalized on interaction even though its appearance score stays high. These choices mitigate rather than quantify copy-paste behavior; disentangling genuine preservation from replication requires multiple references per identity across varied poses and conditions, which we leave to future multi-reference extensions of MIB.

Positive Impacts. MIBE contributes a rigorous, human-aligned diagnostic benchmark (MIB) and evaluator (MIE) that advances transparency and fine-grained error attribution in multi-subject personalized image generation. By decoupling failures into Existence, Appearance, and Interaction rather than reporting a single opaque score, MIBE gives developers actionable signal for reducing hallucinated subjects and identity bleeding in downstream generative systems.

Dataset Impacts. MIB sets a safer standard for benchmark creation: every released image is either generated synthetically or curated under permissive licenses, minimizing the privacy violations, copyright risk, and explicit or harmful content that accompany web-scraped photograph collections.

Dual-Use and Deepfake Risks. A high-fidelity evaluator can also be used to optimize personalized image generators, which could indirectly facilitate more convincing multi-person synthetic media (e.g., deepfakes) if misused. We release MIE as a diagnostic and research tool and encourage downstream users to pair it with existing content-provenance and watermarking safeguards.

Bias and Representation. Both MIB's demographic composition and MIE's pretrained vision-language judges can carry residual biases from their construction and pretraining pipelines respectively. We mitigate this through an explicitly balanced reference pool and dual-judge cross-validation during annotation, but continued human auditing and further diversification of the benchmark remain necessary before deployment in sensitive real-world applications.