WITHEVERYONE: UNIFIED PLANNING AND IDENTITY GROUNDING FOR GROUP IMAGE GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of WithEveryone.** WithEveryone generates coherent group images from five to ten reference identities. A single model reasons about which references participate, plans identity–layout bindings, person regions, and poses, and generates the image conditioned on that plan, so identity and composition are decided in one context rather than by separate modules.

ABSTRACT

Identity-preserving image generation becomes increasingly unreliable when a scene must contain many specified people. Beyond retaining each identity, the model must bind every reference to a distinct person and location, while training-time identity losses must establish correspondence among several noisy predicted faces. We introduce WithEveryone, a unified framework for generating group images of 5–10 reference identities. WithEveryone injects each selected identity as an addressed token, predicts a structured identity–layout plan, and renders the plan as a visual condition. Its key objective, Layout-Grounded ID Loss, uses annotated face regions to supervise the intended identities directly, avoiding unstable embedding-based face matching; ID Representation Forcing additionally trains a prediction for each identity before image synthesis. WithEveryone achieves the highest target-context identity similarity, improving face similarity from 0.462 for GPT-Image 2 to 0.499, while reducing copy-paste artifacts from 0.169 to 0.055. It further covers 97.3% of the requested identities with a duplicate rate of only 2.8%. These results show that explicit identity–layout grounding enables identity-preserving generation to scale to larger groups without relying on direct reference-face copying.

1 INTRODUCTION

Recent image generation models, especially unified multimodal models (UMMs), have advanced rapidly in visual fidelity, text understanding, and controllability, and identity-preserving generation

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.

further lets users specify a person through one or more reference images (Li et al., 2024; Wang et al., 2024c). However, existing open academic methods and benchmarks primarily address one or a small number of people, typically covering one to five reference identities (IDs) (Xu et al., 2025; Cheng et al., 2025; Borse et al., 2025); even strong proprietary systems such as Nano Banana Pro support at most five people (Google DeepMind, 2025). Real group-image creation requires jointly generating more specified people rather than repeatedly generating independent single-person images. We study group-image generation from five to ten reference IDs in one semantically and visually coherent image.

As the number of reference IDs increases, identity similarity is widely reported to decrease (Xu et al., 2025; Kim et al., 2024; Zhang et al., 2025; He et al., 2024), and models tend to duplicate faces or merge IDs in multi-person scenes (Borse et al., 2026a). When more reference images, visual tokens, and identity conditions share one generation context, each identity signal is more easily diluted, and cross-identity interference and incorrect binding become more likely. Supplying reference images or identity embeddings as conditions therefore does not guarantee that the model keeps reading every identity during generation: large-group generation needs a mechanism that both provides each reference identity and constrains the context to retain its representation.

Prior work also shows that conditioning alone is not enough, and that supervising the generated face directly with an identity loss is necessary for strong identity preservation (Guo et al., 2024; Xu et al., 2025). Such losses, however, are formulated for a single person or at most two. Extending them to a large group is not a matter of running the same loss more times, because the loss must first decide which generated face belongs to which reference, and existing formulations recover that correspondence by matching face embeddings in the generated image. Early in training or at high noise levels, however, the predicted faces are nearly identical, so this matching becomes unreliable and the identity loss fails or even becomes counterproductive.

More references also complicate spatial and bodily relationships. Identity similarity answers who is generated, but not where people appear, what poses they take, or how they are organized. As the group grows, pairwise relations multiply and one misplaced person propagates errors through the surrounding composition, making overlaps, limb errors, and crowded layouts especially likely. We therefore offload this planning burden to the understanding side of a unified model before image generation.

To address these challenges, we propose WithEveryone, a UMM framework shown in Figure 1. On the identity side, we extract an identity embedding for each reference person, project it into an ID token, and bind that token to the planned person it should control, so that references form an addressed set rather than an unordered pool. Because ID tokens are easily ignored in long multi-ID sequences, *ID Representation Forcing* adapts representation forcing (Wang et al., 2026) to identities, requiring the model to predict a representation aligned with each target person before synthesis. At the output level, we introduce a flow-compatible *Layout-Grounded ID Loss* (LG-ID Loss), which removes the matching problem described above rather than approximating it: the correspondence is read off the layout annotation, so each identity is supervised in its own annotated region no matter how many people share the image.

On the understanding side, the model autoregressively predicts a structured *Layout Chain of Thought* (Layout CoT) containing person indices, bounding boxes, identity–layout bindings, and pose keypoints, extending ATLAS-style planning (Liu et al., 2026) to an identity-aware setting in which every planned person is tied to a reference identity; a deterministic renderer then converts the plan into a visual condition. ID tokens and Representation Forcing determine who to generate, Layout CoT determines where and how each person appears, and identity–layout binding joins the two in a single plan.

We evaluate WithEveryone on a five-to-ten-person benchmark against academic, open-source, and proprietary systems, including GPT-Image 2 and the Nano Banana and Seedream families. WithEveryone reaches the highest similarity, with a Sim(Tgt) score of 0.499 against 0.462 for GPT-Image 2. Its copy-paste artifact score (Xu et al., 2025) of 0.055 against 0.169 suggests that this similarity relies less on copying reference faces. These results show a better overall balance among identity, copying, and image quality than any single system we compare against. Our contributions are threefold:

- We propose interleaved identity and layout reasoning inside a single unified model: each selected reference is loaded as an ID token, a structured Layout CoT extends ATLAS-style

planning with identity–layout binding and joint prediction of face regions, body extents, and poses, and a deterministic renderer turns the plan into a visual condition. ID Representation Forcing adapts representation forcing by adding supervised per-identity predictions before image synthesis, encouraging the selected identities to remain addressable in the shared context.

- We propose the LG-ID Loss, which grounds output-side identity supervision in the layout annotation instead of in embedding-based face matching. This removes the failure mode that prevents existing identity losses from being applied beyond two people, and it is the single largest source of identity gain in our system.
- We build and systematically evaluate a group-image generation system for five to ten people. On an identity-disjoint benchmark it achieves higher target-context identity similarity than GPT-Image 2 (Sim(Tgt) 0.499 versus 0.462) at a lower copy-paste artifact (0.055 versus 0.169), and our analysis separates plan prediction from plan execution as distinct sources of the remaining error.

2 Related Work

Identity-Preserving and Identity-Consistent Generation. Identity-preserving generation has progressed from single-person methods (Wang et al., 2024c; Li et al., 2024; Valevski et al., 2023; Wang et al., 2024b; Yan et al., 2023; Xiao et al., 2025) to multi-person customization (Zhang et al., 2025; He et al., 2024; Kim et al., 2024; Cheng et al., 2025). Existing methods, however, remain limited to small groups and can trade identity similarity for copy-paste behavior (Chen et al., 2025; Mou et al., 2025; Jiang et al., 2025; Xu et al., 2025); reported examples from WithAnyone and UMO include at most four and three people, respectively. Output-side identity supervision is similarly limited: PuLID (Guo et al., 2024) handles one identity, while WithAnyone uses embedding-based Hungarian matching for at most two. Under training-time noise, this matching becomes unreliable as the number of faces grows. Our LG-ID Loss instead obtains correspondence from layout annotations, enabling output-side supervision for five to ten people.

Text-Image-to-Image Models and Unified Models. Reference-conditioned generation is now supported by open models (Batifol et al., 2025; Wu et al., 2025a; Team et al., 2025; Diao et al., 2026; Cai et al., 2026) and proprietary systems (Google DeepMind, 2026; 2025; ByteDance Seed Team, 2026; OpenAI, 2025; 2026). Unified architectures such as HunyuanImage 3 and BAGEL generate text and images in a shared context, enabling in-context planning before synthesis (Cao et al., 2025; Deng et al., 2025; Zhou et al., 2025).

Layout Planning and Generation. Most layout-conditioned methods use a separate planner and image generator (Feng et al., 2023; Gupta & Kembhavi, 2023; Yang et al., 2024; Zhang et al., 2024; Fang et al., 2025; Duan et al., 2025; Borse et al., 2026b), whereas ATLAS and PlanGen unify planning and synthesis in one contextual sequence (Liu et al., 2026; He et al., 2025). We extend this setting from text-to-image to identity-conditioned generation: each planned person is bound to a reference identity, with jointly predicted face regions, body extents, and poses rendered as a visual condition.

3 Method

3.1 Overview

We consider group-image generation from a text prompt and a set of five to ten reference identities. Our model follows a transfusion-style (Zhou et al., 2025) mixture-of-transformers multimodal architecture (Deng et al., 2025; Cao et al., 2025) in which text and structured reasoning are predicted autoregressively, while target-image latents are learned through flow matching (Lipman et al., 2022). This shared causal context lets high-level decisions made before image generation directly condition the subsequent image tokens.

As shown in Figure 2, WithEveryone prepares three complementary conditions before synthesizing the target image: it selects the reference identities participating in the scene and loads a compact representation for each, predicts a structured multi-person layout together with the association

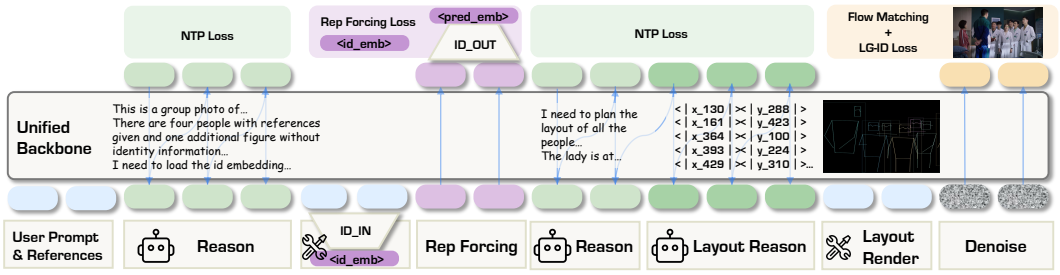


Figure 2: Overview of WithEveryone. The model loads selected reference identities as ID tokens, predicts structured Layout CoT, renders a visual layout condition, and predicts identity representations before generating the target group image. Modality-boundary tokens such as "BOI" are omitted, and the actual CoT order follows the text.

between identities and planned people, and finally predicts target identity representations from the accumulated context. The rendered layout condition and the representation scaffold jointly condition flow-based image generation.

3.2 Interleaved Identity and Layout Reasoning

ID binding and Representation Forcing. The model selects requested reference people while allowing unreferenced people, and complements low-level image features with a high-level identity signal, analogous to ViT/VAE features (Dosovitskiy et al., 2020): each selected identity's 512-dimensional ArcFace embedding (Deng et al., 2019; 2020) is projected through an MLP into an ID token and bound in the subsequent layout. Because these sparse tokens can be diluted in a 10–20K-token context, ID Representation Forcing (Wang et al., 2026) places one prediction token per target-linked identity before image synthesis and maps its contextual hidden state $\mathbf{h}_i^{\text{rep}}$ back to the ArcFace space:

$$\hat{\mathbf{e}}_i = g_{\text{out}}(\mathbf{h}_i^{\text{rep}}), \qquad \mathcal{L}_{\text{RF}} = \frac{1}{M}\sum_{i=1}^{M}\left(1 - \cos\left(\hat{\mathbf{e}}_i, \mathbf{e}_i^{\text{tgt}}\right)\right), \tag{1}$$

where g_{out} is an MLP head, $\mathbf{e}_i^{\text{tgt}}$ is the target-person embedding, and M is the number of supervised identities. Unlike next-token prediction or flow matching, this loss directly aligns a continuous identity prediction from the shared context; the resulting scaffold stays causally available to later image tokens (Appendix C.2).

Structured Layout CoT. ID-preserving generation must also resolve the spatial and bodily relationships among people, which we formulate as a structured Layout CoT. Following ATLAS (Liu et al., 2026), the model autoregressively predicts identity–layout bindings, person and face regions, body extents, and pose keypoints using a discretized coordinate vocabulary with 2,002 tokens, corresponding to 1,001 positions on each axis. The stages are emitted in a fixed causal order, so each decision is conditioned on the identities and regions already committed; short natural-language connectors carry the scene state across stages, while the spatial fields remain parseable and directly supervised. The model then calls a rendering tool that draws the planned face regions, body regions, and pose skeletons onto a blank canvas at the target aspect ratio, which is inserted into the context as a condition image.

3.3 Training Data and Objectives

Training data and sequence. Each training sample interleaves reference images and the prompt, identity selection and ID loading, layout planning, the rendered layout condition, a recaption, the identity prediction, and the target image, in that causal order; Figure 3 shows how these signals are derived from real group photographs, and Appendix C.1 gives the full sequence and annotation pipeline.

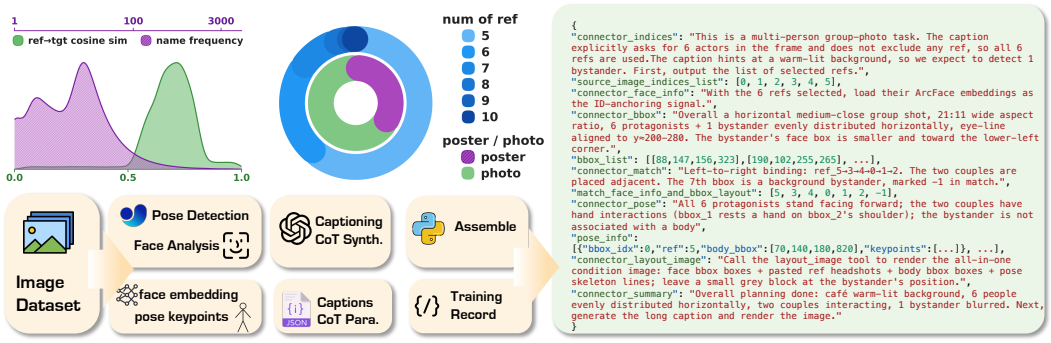


Figure 3: **Data statistics, pipeline, and example.** We report the distributions of identity occurrence frequency, reference count, and real photographs versus collage and poster images; the data-construction pipeline; and an example training sample. The corpus contains 400K group-image samples, and the reference-count distribution covers the five-to-ten-person range used at evaluation time.

Layout-Grounded ID Loss. Existing output-side identity losses (Guo et al., 2024; Xu et al., 2025) either handle one person or recover correspondence by matching generated-face embeddings, which becomes unstable when noisy one-step predictions contain five to ten barely distinguishable faces. PuLID avoids correspondence by supervising one identity, whereas WithAnyone uses Hungarian matching and reports it for at most two. At high noise, arbitrary pairings pull one identity toward another, cancelling rather than sharpening supervision. We remove the matching problem rather than improving the matcher: we take each identity's annotated face box from the target image and crop the predicted image at the same box, so every crop pair is correct by construction for ArcFace supervision (Deng et al., 2019). This holds only at sufficiently low noise, where generated faces are recognizable and follow their boxes; a timestep threshold enforces this (Appendix B.6), and plan adherence converges before identity preservation (Section 4.4). Because correspondence no longer depends on generated appearance, **Layout-Grounded ID Loss** (LG-ID Loss) scales independently of group size and provides the largest individual identity gain in our ablation (Section 4.3).

Joint objective. Different parts of the sequence receive supervision appropriate to their representation. Connector text, structured layout tokens, summaries, and recaptions use next-token prediction; target-image latents use flow matching; and identity predictions use the cosine alignment above. Given a noisy latent $\mathbf{x}_t$ and predicted velocity $\mathbf{v}_\theta(\mathbf{x}_t, t)$ under the reverse-flow convention, the one-step clean-image estimate used by the LG-ID Loss is

$$\hat{\mathbf{x}}_{\text{clean}} = \mathbf{x}_t - t\,\mathbf{v}_\theta(\mathbf{x}_t, t). \tag{2}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{NTP}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}} + \lambda_{\text{RF}}\mathcal{L}_{\text{RF}} + \lambda_{\text{ID}}\mathcal{L}_{\text{ID}}. \tag{3}$$

We set $\lambda_{\text{FM}} = 1.0$, $\lambda_{\text{RF}} = 1.0$, and $\lambda_{\text{ID}} = 0.5$; Appendix C.3 lists the backbone, initialization, data scale, optimizer, and hardware. At inference time, the model predicts identity selection, Layout CoT, and recaption autoregressively; the renderer deterministically constructs the layout condition from the predicted plan. Target identity embeddings are used only as training targets for representation forcing, not as inference-time conditions.

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate on 210 real group images with five to ten reference identities, stratified into 60, 50, 40, 30, 20, and 10 examples, respectively. Each target contains exactly one detected face per reference and no duplicate identities. All benchmark identities and any training images containing them are held out, so the comparison tests unseen people rather than image-level memorization.

Table 1: **Quantitative comparison on our 5–10-person benchmark.** All models are scored on the same 210 examples, and the identity similarities are averaged over the three face encoders of the main protocol. Coverage is the fraction of references whose best-matching generated face reaches a similarity of 0.20, and Dup the fraction that collapse onto a face already claimed by another identity. , , and mark the first-, second-, and third-best results; for Copy-Paste only the three methods with the highest Sim(Ref) are ranked, since lower similarity means less copying naturally.

Method	Identity Similarity			Identity Coverage		Generation Quality		
	Sim(Tgt) ↑	Sim(Ref) ↑	Copy-Paste ↓	Coverage ↑	Dup ↓	CLIP-I ↑	DINO-I ↑	CLIP-T ↑
Academic identity-preserving methods								
WithAnyone	0.405	0.483	0.096	0.957	0.045	0.807	0.695	0.281
UMO	0.371	0.484	0.112	0.630	0.258	0.780	0.663	0.286
UniPortrait	0.339	0.464	0.115	0.635	0.187	0.679	0.415	0.301
DreamO	0.297	0.331	0.027	0.298	0.299	0.724	0.611	0.282
ID-Patch	0.225	0.259	0.032	0.247	0.224	0.609	0.320	0.331
Open-source general-purpose models								
FLUX.2 Klein	0.264	0.265	0.002	0.314	0.265	0.787	0.685	0.291
Qwen-Image-Edit	0.253	0.289	0.018	0.176	0.291	0.614	0.312	0.227
LongCat-Image-Edit	0.287	0.288	−0.001	0.381	0.264	0.802	0.657	0.288
OmniGen2	0.267	0.275	−0.002	0.418	0.267	0.774	0.633	0.288
SenseNova U1	0.253	0.243	−0.009	0.213	0.256	0.797	0.686	0.289
BAGEL	0.223	0.225	0.003	0.136	0.276	0.752	0.635	0.282
HiDream-O1	0.353	0.376	0.026	0.780	0.190	0.806	0.692	0.286
Proprietary systems								
Nano Banana Pro	0.453	0.478	0.041	0.674	0.148	0.839	0.712	0.275
Nano Banana 2	0.451	0.480	0.045	0.884	0.099	0.860	0.731	0.276
GPT-Image 2	0.462	0.583	0.169	0.905	0.075	0.853	0.719	0.270
Seedream 4.5	0.407	0.495	0.114	0.859	0.175	0.829	0.698	0.286
Seedream 5.0 Pro	0.436	0.522	0.114	0.913	0.065	0.850	0.715	0.275
WithEveryone	0.499	0.540	0.055	0.973	0.028	0.861	0.716	0.273

Baselines and protocol. We compare academic identity methods (UMO (Cheng et al., 2025), WithAnyone (Xu et al., 2025), UniPortrait (He et al., 2024), DreamO (Mou et al., 2025), and ID-Patch (Zhang et al., 2025)), open models (FLUX.2 Klein (Labs, 2025a), Qwen-Image-Edit (Wu et al., 2025a), LongCat-Image-Edit (Team et al., 2025), OmniGen2 (Wu et al., 2025b), SenseNova U1 (Diao et al., 2026), BAGEL (Deng et al., 2025), and HiDream-O1 (Cai et al., 2026)), and proprietary Nano Banana (Google DeepMind, 2025; 2026), Seedream (ByteDance Seed Team, 2026), and GPT-Image 2 (OpenAI, 2026). Every system runs on all examples at its highest supported resolution; WithEveryone uses 2K for Table 1 and 1K otherwise. ID-Patch and WithAnyone receive ground-truth layouts because they do not plan them.

Metrics. Following WithAnyone (Xu et al., 2025), $\mathrm{Sim(Ref)}$ and $\mathrm{Sim(Tgt)}$ average ArcFace (buffalo_l) (Deng et al., 2019), FaceNet (Schroff et al., 2015), and AdaFace (Kim et al., 2022); ablations use ArcFace only. Assigning generated faces to references is reliable on the clean final image; the LG-ID Loss instead addresses noisy one-step training predictions. Human-validated Copy-Paste (Xu et al., 2025) measures reuse; Coverage is the fraction of references matched above 0.20; and Dup is the fraction whose best generated face is also selected by another reference. DINO-I (Oquab et al., 2023) and CLIP-I/T (Radford et al., 2021) measure image/text alignment; Count is count accuracy; and Plan IoU is mean matched IoU between supplied face boxes and detected generated faces, measuring own-plan or oracle-plan execution. Appendix C gives full definitions.

4.2 Cross-Model Comparison

Identity, coverage, and copying. In Table 1, WithEveryone has the highest Sim(Tgt) (0.499), second-highest Sim(Ref) (0.540), best Coverage/Dup pair (0.973/0.028), and lower Copy-Paste than GPT-Image 2 (0.055 versus 0.169), indicating less rigid copying than the strongest baseline (Figure 5). Its paired Sim(Tgt) margin over GPT-Image 2 is $+0.038$ with a 95% bootstrap interval of $[+0.027, +0.048]$, favoring WithEveryone on 73% of examples (Appendix B.1). Seedream 5.0 Pro reaches 0.913/0.065 and GPT-Image 2 0.905/0.075, so WithEveryone rarely omits a person or

collapses two references onto one face. Open models mostly stay below 0.42 Coverage, editing prompt-implied people rather than composing the full group. A secondary one-encoder score is 0.614 versus 0.566 for GPT-Image 2; because it uses a separate scoring pipeline, conclusions follow the three-encoder table.

Generation quality and scaling. WithEveryone has the highest CLIP-I (0.861, statistically tied with Nano Banana 2 at 0.860; Appendix B.1) and competitive DINO-I (0.716). CLIP-T is lower (0.273 versus 0.331/0.301 for ID-Patch/UniPortrait), consistent with text alignment being easier to satisfy when the generated people need not match specific references. Across five to ten references (Figure 4), WithEveryone ranks first at every group size and declines from 0.629 to 0.571, versus 0.593 to 0.496 for GPT-Image 2 and 0.412 to 0.330 for UMO; its distribution is tighter, while GPT-Image 2 has heavier tails. Because each group size has few examples, we read this as an aggregate trend rather than a per-size claim (Appendix B.2).

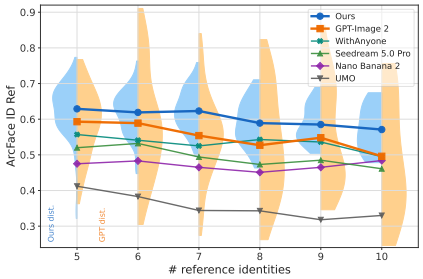


Figure 4: **Scaling to more references.** Lines show mean ArcFace similarity to the references; distributions are shown for WithEveryone and GPT-Image 2.

Table 2: **Ablation results.** The rows form three incremental branches from the same base: layout (P1–P3), ID token and Representation Forcing (P1, P4, P5), and LG-ID Loss (P1, P6); P7 is the complete configuration and additionally uses the text-to-layout corpus.

	Variant	Sim(Ref) ↑	Sim(Tgt) ↑	Count ↑	Coverage ↑
P1	Default	0.339	0.304	0.771	0.741
P2	Model Layout	0.364	0.316	0.828	0.813
P3	GT Layout	0.412	0.367	**0.958**	0.891
P4	ID Token	0.351	0.313	0.827	0.761
P5	ID Token + RF	0.364	0.328	0.817	0.782
P6	LG-ID Loss only	0.506	0.435	0.845	0.947
P7	Full Model	**0.555**	**0.461**	0.869	**0.960**

4.3 Ablation Studies

Table 2 evaluates layout conditioning (P1–P3), ID tokens and Representation Forcing (P1, P4, P5), LG-ID Loss (P1, P6), and the full model (P7). Variants are trained independently under the same recipe, except that P7 also uses the text-to-layout corpus (Appendix C.1); Appendix B.8 gives the full setup of each branch.

Effect of the LG-ID Loss. LG-ID Loss provides the largest individual gain, raising Sim(Ref)/Sim(Tgt) from 0.339/0.304 to 0.506/0.435 and Count/Coverage from 0.771/0.741 to 0.845/0.947. Because each identity is supervised inside its annotated region, the loss is reduced only by placing the right person in the right location; it therefore carries spatial as well as identity information, so the layout-only gain below understates spatial grounding. P7 combines LG-ID Loss, ID tokens, Representation Forcing, Layout CoT, and text-to-layout data, reaching 0.555/0.461; these simultaneous changes prevent attribution of the margin. Its Coverage is not a threshold artifact: raising the matching threshold from 0.20 to 0.25, 0.30, and 0.35 lowers Coverage by only 0.010, 0.017, and 0.027, while just 3.6% of references lie near the boundary (Appendix C.7). Lowering the threshold to 0.15 changes Coverage by only $+0.012$, further confirming robustness.

Effect of Layout CoT. P1 omits Layout CoT and ID tokens, while P2 and P3 use model-predicted and target-derived layouts, respectively. Model-predicted layout raises Sim(Ref) from 0.339 to 0.364, Count from 0.771 to 0.828, and Coverage from 0.741 to 0.813, with little Sim(Tgt) change. Ground-truth layout further raises Sim(Ref)/Sim(Tgt) to 0.412/0.367 and improves every layout metric. As a target-derived oracle rather than an achievable planning gain, it shows that **high-quality layout improves both composition and identity** and that planning quality remains a bottleneck.

Effect of ID tokens and Representation Forcing. ID tokens yield small, consistent gains, with the largest increase in Count (0.771 to 0.827); Representation Forcing further improves



Figure 5: **Qualitative comparison with proprietary models.** ■, ■, and ■ mark faces that reproduce a person already generated elsewhere in the image, that are not recognizable as their bound reference, and that appear transplanted without adapting to the pose, lighting, or viewpoint of the scene. These author annotations are illustrative; their quantitative counterparts are the duplicate rate, Sim(Ref), and Copy-Paste.

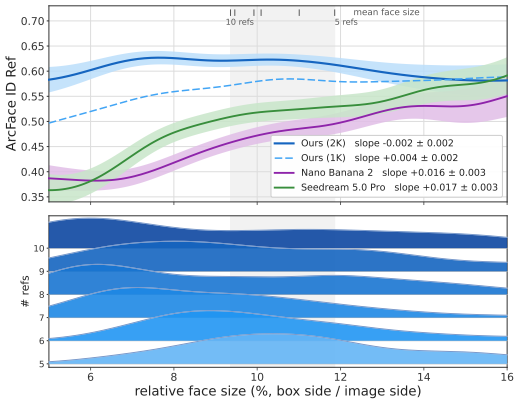
Sim(Ref)/Sim(Tgt) by $+0.013/+0.015$. Together with the identity-specific attention in Figure 8, these modest gains support improved identity addressability rather than a large direct identity gain.

4.4 Analysis

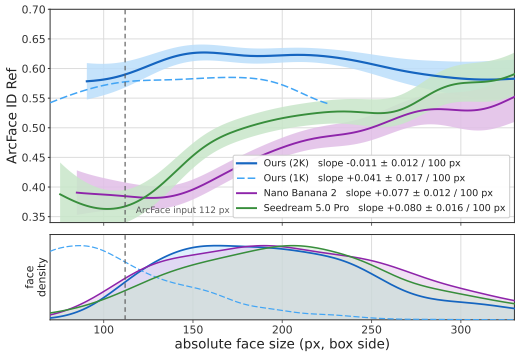
We diagnose scaling through face resolution, plan execution, and identity routing.

Planning versus execution. Following its own plan, the full model converges to 0.814 Plan IoU in Figure 7, matching the flat 0.79–0.80 reference of the oracle-plan variant, so it executes its own plan about as faithfully as a supplied one. Substituting that ground-truth layout nevertheless improves identity and coverage (Table 2), locating the residual error in plan prediction rather than execution. Since execution converges before identity preservation (Figure 12), the main opportunity lies on the understanding side; Appendix C.8 defines the Relative Layout Score (RLS), which scores plan quality against the ground-truth configuration, and Appendices B.4 and B.6.3 report it.

Representation Forcing mechanism. Figure 8 shows layer-24 diagonal attention of 0.378–0.574 versus 0.104–0.121 off-diagonal, a $+0.275$ to $+0.467$ identity-specificity gap. This supports identity-specific routing at supervised prediction positions, but not a causal effect without a no-RF control or comparison to generated face patches.



(a) Relative face size, as a fraction of the image side; the grey band spans the mean face size from ten references on its left edge to five on its right, the entire range that group size produces.



(b) Absolute face size in pixels; the dashed line marks 112 px, the input resolution of the ArcFace encoder, below which a face has to be upsampled before it can be compared.

Figure 6: **Identity similarity against face size.** Curves are local means with 95% confidence bands, annotated with the fitted slope. The lower panel of (a) shows the distribution of relative face size for each group size, and the lower panel of (b) shows the face-size distribution of each system.

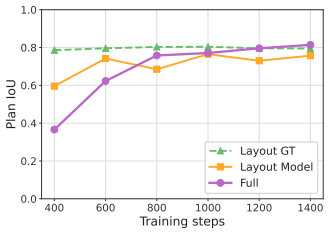


Figure 7: **Plan IoU convergence.** The model rapidly converges to follow the planned layout.

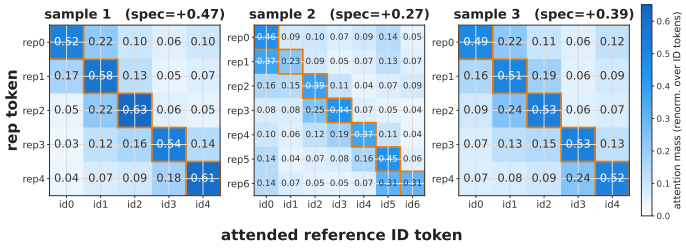


Figure 8: **Attention at Representation Forcing prediction positions.** In the three illustrated examples, layer-24 attention from the supervised identity predictions to reference ID tokens shows diagonal dominance.

Face size and scaling degradation. Similarity could fall with group size simply because faces get smaller. Figure 6 shows that the baselines depend strongly on face size, gaining 0.016–0.019 similarity per percentage point of relative face side, whereas our slope is nearly flat at -0.002. Controlling for face size removes most of the group-size slope for GPT-Image 2 and Seedream 5.0 Pro but leaves ours essentially unchanged. The reduction in face resolution therefore does not explain our remaining group-size degradation, which may reflect other factors that grow with group size, including cross-identity interference, occlusion, and compositional complexity. Within a fixed pixel band WithEveryone leads below 200 px and GPT-Image 2 overtakes it above. This pattern is qualitatively consistent with GPT-Image 2's higher Copy-Paste, although the association does not identify a causal explanation. Appendix B.3 gives the fits and the per-band numbers.

5 Conclusion

We presented WithEveryone, combining ID tokens, Representation Forcing, identity-aware Layout CoT, and LG-ID Loss for five-to-ten-person generation. On an identity-disjoint benchmark, it achieves the best Sim(Tgt) (0.499 versus 0.462) and coverage with lower Copy-Paste than GPT-Image 2 (0.055 versus 0.169). LG-ID Loss provides the largest gain, Layout CoT adds coverage and spatial control, and diagnostics identify plan prediction as the remaining bottleneck within this benchmark. More broadly, multi-person identity supervision benefits from explicit spatial addresses. Appendix E discusses limitations and responsible use.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing the manuscript and with portions of software coding. Following common practice, the short captions and detailed recaptions of our training images were generated by a vision-language model and serve only as textual annotations. No synthetic images were used: all training and benchmark images are collected from public sources rather than generated by a model. Generative AI tools were not used to develop theoretical models or mathematical claims, write proofs, propose hypotheses, design or provide feedback on the research methodology or experiments, perform qualitative or thematic data analysis, or interpret experimental results. All AI-assisted text and code were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We constructed the training dataset from images openly posted on public websites, including news outlets, official social-media accounts, and other publicly accessible pages, and did not collect from private or access-restricted sources. To reduce potential harm, we excluded images of minors and intimate, medical, or otherwise sensitive situations. The dataset is intended solely for non-commercial academic research, and any access to the full dataset will be governed by a research-only license and accompanying usage guidelines. Identity-preserving generation, including that offered by the open and proprietary systems we compare against, can be misused to depict real people without their consent or to impersonate them; this risk is shared by the task in general rather than specific to our method. We recommend that any deployment obtain consent from the referenced people and disclose generated content through provenance signals such as watermarking. Appendix E further discusses responsible use.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the model design, objectives, benchmark protocol, baselines, and metrics. Appendix C documents data construction, training configuration, metric definitions, thresholds, and additional ablations. The in-house base model and training corpus limit exact end-to-end reproduction outside the authors' environment; the reported details are intended to support reimplementation and evaluation with comparable resources.

REFERENCES

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Shubhankar Borse, Seokeon Choi, Sunghyun Park, Jeongho Kim, Shreya Kadambi, Risheek Garrepalli, Sungrack Yun, Munawar Hayat, and Fatih Porikli. Multihuman-testbench: Benchmarking image generation for multiple humans. *arXiv preprint arXiv:2506.20879*, 2025.

Shubhankar Borse, Farzad Farhadzadeh, Munawar Hayat, and Fatih Porikli. Resolving the identity crisis in text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36703–36712, 2026a.

Shubhankar Borse, Phuc Pham, Farzad Farhadzadeh, Seokeon Choi, Phong Nguyen, Anh Tran, Sungrack Yun, Munawar Hayat, and Fatih Porikli. Ar2can: An architect and an artist leveraging a canvas for multi-human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 550–560, 2026b.

ByteDance Seed Team. Beyond generation, it understands design: Introducing seedream 5.0 pro, July 2026. URL `https://seed.bytedance.com/en/blog/beyond-generation-it-understands-design-introducing-seedream-5-0-pro`.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

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. *arXiv preprint arXiv:2506.21416*, 2025.

Yufeng Cheng, Wenxu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025.

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.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. *arXiv preprint arXiv:2505.17022*, 2025.

Rongyao Fang, Chengqi Duan, Kun Wang, Linjiang Huang, Hao Li, Shilin Yan, Hao Tian, Xingyu Zeng, Rui Zhao, Jifeng Dai, et al. Got: Unleashing reasoning capability of multimodal large language model for visual generation and editing. *arXiv preprint arXiv:2503.10639*, 2025.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36:18225–18250, 2023.

Google DeepMind. Introducing nano banana pro, November 2025. URL `https://blog.google/innovation-and-ai/products/nano-banana-pro/`.

Google DeepMind. Nano banana 2: Combining pro capabilities with lightning-fast speed, February 2026. URL `https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/`.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In *Advances in Neural Information Processing Systems*, 2024.

Zinan Guo, Pengze Zhang, Yanze Wu, Chong Mou, Songtao Zhao, and Qian He. Musar: Exploring multi-subject customization from single-subject dataset via attention routing. *arXiv preprint arXiv:2505.02823*, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14953–14962, 2023.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024.

Runze He, Bo Cheng, Yuhang Ma, Qingxiang Jia, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Liebucha Wu, Dawei Leng, and Yuhui Yin. Plangen: Towards unified layout planning and image generation in auto-regressive vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18143–18154, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *Advances in Neural Information Processing Systems*, 38:39856–39890, 2026.

Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.

Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024.

Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025a.

Black Forest Labs. Flux.1 krea. `https://huggingface.co/black-forest-labs/FLUX.1-Krea-dev`, 2025b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.

Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025a.

Junhao Liu, Jian-Wei Zhang, Tao Huang, Miles Yang, Zhao Zhong, and Liefeng Bo. Think, plan, paint: Layout-aware reasoning for controllable image generation in unified models. *arXiv preprint arXiv:2607.16409*, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Chong Mou, Yanze Wu, Wenxu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Addendum to gpt-4o system card: Native image generation, 2025. URL https://cdn.openai.com/11998be9-5319-4302-bfbf-1167e093f1fb/Native_Image_Generation_System_Card.pdf.

OpenAI. Introducing chatgpt images 2.0, April 2026. URL https://openai.com/index/introducing-chatgpt-images-2-0/.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, pp. 241–261. Springer, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27080–27090, 2024.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Ultralytics. Ultralytics yolo11, 2024. URL https://github.com/ultralytics/ultralytics.

Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–10, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024a.

Qinghe Wang, Xu Jia, Xiaomin Li, Taiqing Li, Liqian Ma, Yunzhi Zhuge, and Huchuan Lu. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*, 2024b.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024c.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, et al. Representation forcing for bottleneck-free unified multimodal models. *arXiv preprint arXiv:2605.31604*, 2026.

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, 2025a. URL https://arxiv.org/abs/2508.02324.

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. *arXiv preprint arXiv:2506.18871*, 2025b.

Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. 2025c.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025d.

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.

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. *arXiv preprint arXiv:2409.11340*, 2024.

Hengyuan Xu, Liyao Xiang, Hangyu Ye, Dixi Yao, Pengzhi Chu, and Baochun Li. Permutation equivariance of transformers and its applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5987–5996, 2024.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Gang Yu, Xinjun Ma, and Yu-Gang Jiang. Withanyone: Towards controllable and id-consistent image generation. *arXiv preprint arxiv:2510.14975*, 2025.

Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023.

Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Icml*, volume 3, pp. 7, 2024.

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.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Kai-Ni Wang, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, et al. Realcompo: Balancing realism and compositionality improves text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 37:96963–96992, 2024.

Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. Id-patch: Robust id association for group photo personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Table 3: **Dispersion of the main comparison.** Per-example means on the 210 benchmark examples with their standard errors. The last column is the single-encoder ArcFace similarity to the references used in Section 4.2.

Method	Sim(Tgt) ↑	Sim(Ref) ↑	CLIP-I ↑	ArcFace Sim(Ref) ↑
WithEveryone	0.499 ± 0.004	0.540 ± 0.004	0.861 ± 0.004	0.614 ± 0.005
GPT-Image 2	0.462 ± 0.004	0.583 ± 0.008	0.853 ± 0.004	0.566 ± 0.009
Nano Banana 2	0.451 ± 0.005	0.480 ± 0.007	0.860 ± 0.004	0.471 ± 0.008
Seedream 5.0 Pro	0.436 ± 0.005	0.522 ± 0.008	0.850 ± 0.004	0.506 ± 0.009

A Additional Qualitative Results

Figures 9–11 collect a small selection of group images generated by WithEveryone. A larger anonymous gallery, in which each generated group can be inspected together with its reference identities, is available at `https://witheveryone.github.io/`.

B Additional Experiments and Analysis

B.1 Statistical Reliability of the Main Comparison

Table 1 reports means over 210 examples. To show how precisely those means are determined, Table 3 recomputes them per example and gives the standard error for WithEveryone and the three strongest baselines. The identity margins discussed in Section 4.2 are large relative to this dispersion, while the two near-ties are not, and paired comparisons over the same 210 examples, with none excluded for any model, confirm both readings. Our Sim(Tgt) advantage over GPT-Image 2 is $+0.038$ with a 20,000-sample bootstrap interval of $[+0.027, +0.048]$, a Wilcoxon p of 3.8×10^{-11}, and a higher score on 73% of examples. Our CLIP-I margin over Nano Banana 2 is $+0.001$ with an interval of $[-0.006, +0.009]$, a p of 0.98, and a higher score on 48% of examples: the two systems are statistically indistinguishable on this metric, and the best-result mark in that column should not be read as a lead. Against Seedream 5.0 Pro, Sim(Ref) improves by $+0.018$ ($[+0.005, +0.039]$) while Copy-Paste falls by 0.059 ($[-0.085, -0.041]$) and ArcFace similarity rises by $+0.108$ ($[+0.092, +0.131]$), so both directions hold simultaneously.

B.2 Extended Reading of the Main Comparison

This section records the per-metric reading of Table 1 that Section 4.2 states only briefly. The highest Sim(Tgt) of 0.499 indicates that the faces generated by WithEveryone best match the identities as they appear in the target context. Its Sim(Ref) of 0.540 is second only to GPT-Image 2 at 0.583 and higher than both Nano Banana generations at 0.478 and 0.480. The lower Copy-Paste relative to GPT-Image 2, 0.055 versus 0.169, therefore comes with comparable reference similarity and higher target-context similarity, so the identity scores of WithEveryone are less consistent with rigidly reproducing the reference face than those of the strongest baseline.

Similarity scores say how well the generated faces match, but not how many of the requested people are generated as separate individuals; Coverage and Dup measure the latter. WithEveryone covers 0.973 of the reference identities with the lowest duplicate rate of 0.028, ahead of Seedream 5.0 Pro at 0.913 and 0.065 and GPT-Image 2 at 0.905 and 0.075. Among the open general-purpose models only HiDream-O1 reaches a substantial Coverage of 0.780; the others stay at or below 0.418, tending to edit the few people implied by the prompt instead of composing the full group.

The single-encoder ArcFace similarity of 0.614 for WithEveryone against 0.566 for GPT-Image 2 (Table 3) relies on one encoder and a separate scoring pipeline. We therefore treat it as an ArcFace-specific secondary analysis and base all cross-model conclusions on the three-encoder protocol of Table 1.

CLIP-T is the only generation-quality column in which WithEveryone is clearly behind: its 0.273 compares with 0.331 for ID-Patch and 0.301 for UniPortrait. Both methods score far lower on every identity metric, with Sim(Tgt) of 0.225 and 0.339, which is consistent with text alignment being easier to satisfy when the generated people need not match specific references.



Figure 9: **Group images generated by WithEveryone.** The samples cover different group sizes, scene types, and visual styles.



Figure 10: **Additional group images generated by WithEveryone.**



Figure 11: **Additional group images generated by WithEveryone.**

Table 4: **High-resolution comparison.** The first two rows use the same internal pipeline; the last row uses the main 2K benchmark pipeline.

Train / test	Sim(Ref) ↑	Sim(Tgt) ↑	Layout* ↑	Plan IoU ↑
1K / 1K	0.546	0.460	0.740	0.773
2K / 1K	**0.555**	**0.461**	**0.759**	**0.814**
2K / 2K†	0.614	0.511	–	–

*Layout Score, defined in Appendix C.7. †Main benchmark pipeline; included as a contextual reference only.

In the group-size breakdown of Figure 4, the per-example distribution of WithEveryone is also more concentrated, whereas GPT-Image 2 shows substantially heavier upper and lower tails. Because the groups are small, especially at the upper end with 20 and 10 examples for nine and ten references, we read the figure as a consistent advantage across the whole five-to-ten range and a flatter degradation trend, rather than as a claim about any individual group size.

B.3 Face Size and Scaling Degradation

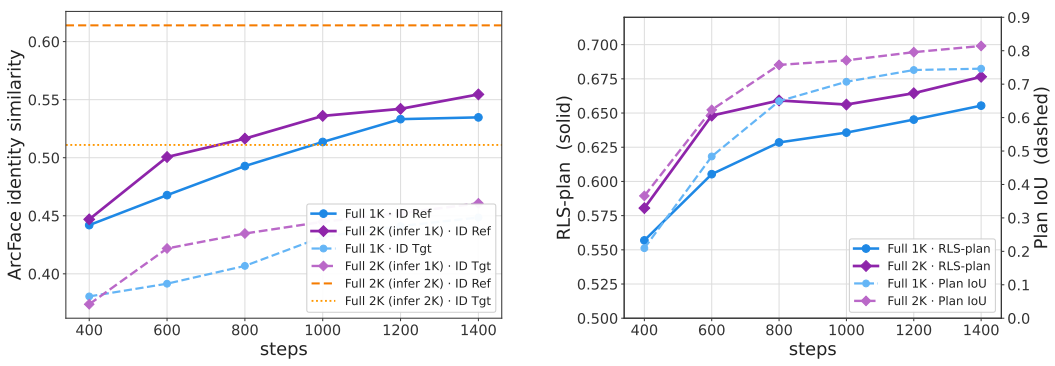
This section expands the analysis summarized in Section 4.4 and refers throughout to Figure 6. For every generated face on the benchmark we record its ArcFace similarity to the reference it is bound to, together with the size of its box, measured both as a fraction of the image side and in pixels. The analysis is per face, and every statement below concerns conditional means. The per-face standard deviation is 0.21–0.24 and is dominated by identity difficulty, pose, and occlusion, so face size accounts for only a small share of the variance of any single face.

Every baseline depends strongly on relative face size, as Figure 6a shows. Nano Banana 2 gains 0.016 ± 0.003 of similarity per percentage point of relative face side, Seedream 5.0 Pro 0.017 ± 0.003, and GPT-Image 2 0.019 ($[+0.015, +0.023]$). WithEveryone at 2K is flat, at -0.002 ± 0.002. This matters because growing the group barely changes how large the faces are: going from five to ten references moves the mean relative face side only from 11.9% to 9.4%. Across that narrow interval, shown in grey in the figure, Nano Banana 2 loses 0.040 of similarity and Seedream 5.0 Pro loses 0.042, while WithEveryone gains 0.006.

We then fit similarity on the number of references across all six group sizes, once with and once without face size as a covariate. For GPT-Image 2 and Seedream 5.0 Pro, controlling for face size removes most of the group-size slope, and what remains has an interval that includes zero; shrinking faces are therefore associated with a substantial part of their degradation. For WithEveryone the slope does not move, staying at -0.014 with an interval that excludes zero. The reduction in face resolution thus does not explain our remaining degradation, but this analysis does not distinguish cross-identity interference from occlusion, compositional complexity, or other factors that vary with group size. We do not quote a numerical share of the face-size association, because the interval on that ratio is too wide to be informative.

Figure 6b repeats the analysis in absolute pixels. The systems differ in output resolution here, so the curves have to be compared within a pixel band rather than along the axis. Within a common band, WithEveryone is the strongest below 200 px, and its margin is largest on the smallest faces: between 112 and 150 px it reaches 0.611, against 0.561 for GPT-Image 2, 0.388 for Seedream 5.0 Pro, and 0.383 for Nano Banana 2. GPT-Image 2 overtakes WithEveryone only in the 200–260 px band, at 0.654 against 0.616, and only 8% of its faces fall there. Our advantage is therefore concentrated on small faces. This pattern is qualitatively consistent with GPT-Image 2's higher Copy-Paste, although it does not establish copying as the cause of the difference.

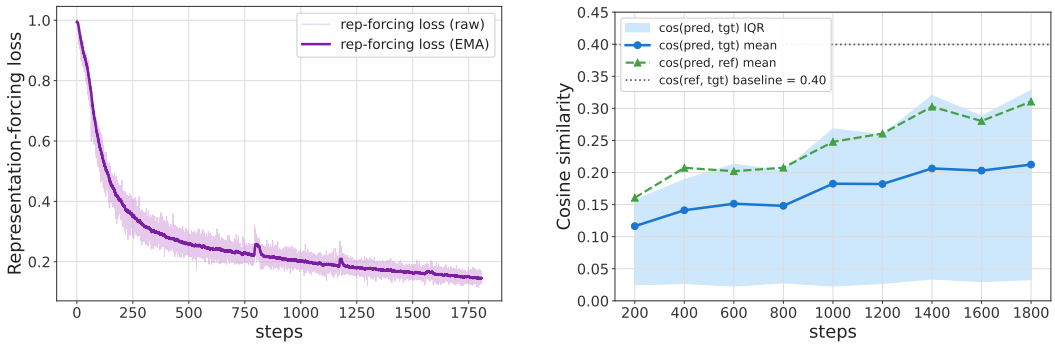
The 1K curve is a reference point rather than a controlled ablation, since it differs from the 2K curve in the reference-image configuration as well as in resolution. It is nonetheless informative about resolution. At 1K, 61% of the faces are smaller than the 112 px input of the ArcFace encoder, against 3% at 2K, so much of the benefit of high-resolution training reported in Appendix B.4 amounts to moving faces out of that regime.



(a) Identity similarity during training.

(b) Planning quality and plan execution.

Figure 12: **High-resolution training.** We compare the Full model trained at 1K and 2K while evaluating both at 1K. The horizontal references in (a) show the final 2K benchmark scores from a different evaluation pipeline.



(a) Representation Forcing training loss.

(b) Inference-time representation alignment.

Figure 13: **Representation Forcing convergence.** The supervised identity prediction is steadily optimized during training and becomes increasingly aligned with both the target and reference identity embeddings.

B.4 High-Resolution Training

Faces in a group image often occupy only a small fraction of the image. Increasing the target resolution therefore allocates more pixels to identity-critical facial regions and may also provide a denser spatial signal for executing the predicted layout. We study this effect by comparing the Full model trained at 1K with the Full model trained at 2K. For a controlled comparison, both models are first evaluated at 1K. We additionally report the final 2K benchmark result to show the operating point used in the main comparison.

Under the controlled 1K evaluation, 2K training improves Sim(Ref) from 0.546 to 0.555 and does not reduce Sim(Tgt) or Layout Score. Its largest gain is in Plan IoU, which rises from 0.773 to 0.814. Figure 12b further shows that the 2K model reaches stronger planning and execution scores throughout the later stage of training. These results indicate that high-resolution training mainly improves how faithfully image generation realizes the spatial plan, while providing a smaller identity benefit at a fixed inference resolution. When the 2K model is evaluated at its native resolution in the main benchmark, it reaches 0.614 Sim(Ref) and 0.511 Sim(Tgt); because that benchmark uses a different scoring pipeline, these values should not be interpreted as a controlled estimate of the gain from changing inference resolution alone.

B.5 Representation Forcing Convergence

Representation Forcing is designed to make the shared context retrieve identity information before image generation. Figure 13a shows that its training objective can be optimized smoothly: the loss decreases from approximately 1.0 to 0.15.

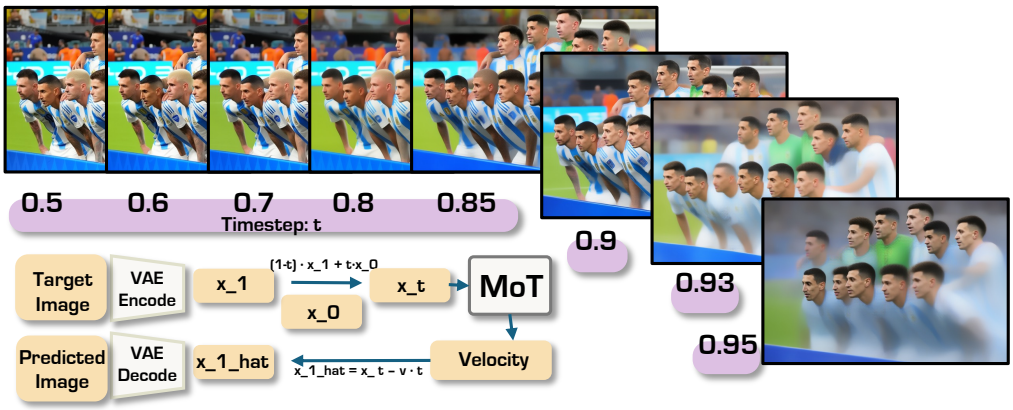


Figure 14: **Training-time one-step prediction and its outputs under different sampled timesteps.** We use $t = 0.85$ as the timestep threshold and compute the LG-ID Loss only when $t \leq 0.85$.

We further evaluate the predicted representation directly at inference time. As shown in Figure 13b, the mean cosine similarity between the prediction and target identity embedding increases from 0.116 at 200 steps to 0.212 at 1800 steps. Its similarity to the reference identity follows the same trend. Both remain below the reference–target baseline of approximately 0.40, so the prediction does not fully reconstruct the target identity embedding. The result supports the narrower conclusion that Representation Forcing progressively aligns the prediction with the intended identity direction, consistent with improved identity addressability.

B.6 LG-ID Loss: Algorithm and Hyper-parameters

B.6.1 One-Step Identity Supervision

The LG-ID Loss augments the image flow-matching objective with identity supervision on a differentiable clean-image estimate. Given a noisy latent $\mathbf{x}_t$ and predicted velocity $\mathbf{v}_\theta(\mathbf{x}_t, t)$, we first compute

$$\hat{\mathbf{x}}_{\text{clean}} = \mathbf{x}_t - t\,\mathbf{v}_\theta(\mathbf{x}_t, t), \tag{4}$$

decode $\hat{\mathbf{x}}_{\text{clean}}$ with the VAE, and compare the resulting faces with their target identities using a frozen ArcFace encoder. Figure 14 summarizes the one-step prediction and visualizes its output at different sampled timesteps.

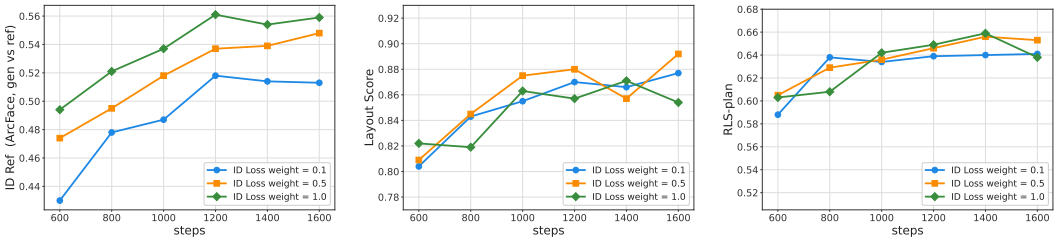
For multi-person supervision, we avoid reordering identities with an embedding-based matching heuristic. The correspondence is already known from the annotation that supervises the Layout CoT: each referenced identity carries a fixed target face bounding box and its landmarks, so cropping the predicted and target images at these regions makes every crop pair correct by construction, and each identity is supervised on its own, independently of how many people share the image. We detect faces in the ground-truth image and match each detection to its dataset bounding box by IoU; bystanders without a referenced identity are ignored. The same ground-truth landmarks then determine corresponding crops in the predicted and target images. The LG-ID Loss is the mean per-face cosine distance, $1 - \cos(\mathbf{e}^{\text{pred}}, \mathbf{e}^{\text{tgt}})$, over valid referenced identities.

B.6.2 Timestep Threshold

At high-noise timesteps, the one-step estimate may not contain a recognizable face, making identity supervision unreliable. As shown in Figure 14, the estimated clean image retains clear identity cues up to $t = 0.85$, whereas facial details degrade rapidly at larger timesteps. We therefore set the hard timestep threshold to 0.85 and compute the LG-ID Loss only when $t \leq 0.85$.

B.6.3 Loss Weight

We vary only the LG-ID Loss weight while keeping the remaining Full-model recipe fixed. As shown in Figure 15a, a larger weight consistently improves identity preservation at both training steps:



(a) Identity preservation. (b) Layout Score. (c) RLS.

Figure 15: **LG-ID Loss weight ablation.** Increasing λ_{ID} substantially improves Sim(Ref), whereas Layout Score and RLS remain in the same range without a stable ordering across weights.

Sim(Ref) at 1200 steps is 0.518, 0.537, and 0.561 for $\lambda_{\text{ID}} = 0.1$, 0.5, and 1.0, respectively; at 1600 steps, the corresponding values are 0.513, 0.548, and 0.559. The main experiments use $\lambda_{\text{ID}} = 0.5$, and this ablation shows that identity preservation can be pushed further by raising the weight.

The gain in identity does not produce a corresponding degradation in planning. Figures 15b and 15c show tightly overlapping trajectories across the three weights for the two composite layout measures, the Layout Score of Appendix C.7 and the Relative Layout Score (RLS) of Appendix C.8: Layout Score remains within approximately 0.80–0.89, and RLS within approximately 0.59–0.66. Larger identity weights can slow early layout improvement slightly, but the difference disappears later in training. Within the tested range, increasing λ_{ID} therefore primarily affects identity preservation and has no visible cost in converged layout quality.

B.7 Effect of the LG-ID Loss

Table 5 expands the main-paper ablation of the LG-ID Loss with the complete set of identity and layout measures, including the Layout Score defined in Appendix C.7. The loss alone raises Sim(Ref) from 0.339 to 0.506 and Sim(Tgt) from 0.304 to 0.435. At the same time, Layout Score increases slightly from 0.690 to 0.700, Count from 0.771 to 0.845, and Coverage from 0.741 to 0.947. Direct identity supervision therefore does not trade layout validity for identity similarity. Because each identity is supervised inside its annotated region, the loss is reduced by putting the right person in the right place, so it carries a spatial signal in addition to an identity one.

Table 5: **Effect of the LG-ID Loss.** It gives the largest individual identity gain while preserving or improving all reported layout measures.

Variant	Sim(Ref) ↑	Sim(Tgt) ↑	Layout ↑	Count ↑	Coverage ↑
Default	0.339	0.304	0.690	0.771	0.741
+ ID Loss	0.506	0.435	0.700	0.845	0.947
Full	**0.555**	**0.461**	**0.759**	**0.869**	**0.960**

The complete configuration combines the LG-ID Loss with ID tokens, Representation Forcing, Layout CoT, and additional text-to-layout supervision, raising Sim(Ref) and Sim(Tgt) to 0.555 and 0.461, respectively, while reaching a Layout Score of 0.759. This result improves over the LG-ID-only branch, but because several components and the training corpus change together, the comparison does not isolate the source of the additional gain.

B.8 Setup and Reading of the Ablation Branches

All variants in Table 2 start from the same base model and are trained independently under the same recipe, with sample-wise rather than fixed training resolutions, and are evaluated at 1K with ArcFace-only identity scoring. P1 is the default model, trained on the same dataset but without Layout CoT or ID tokens. P2 and P3 are both trained with Layout CoT; at inference, P2 uses the layout the

model predicts itself, whereas P3 replaces it with the ground-truth layout. Because the P3 layout is derived from the target image, P3 is an oracle upper bound rather than an estimate of an achievable planning gain. It establishes that high-quality layout conditions can improve both composition and identity preservation, while the gap between P2 and P3 identifies planning quality as an important remaining source of error.

P7 additionally trains on the text-to-layout corpus of Appendix C.1 and scores higher than P2 on both identity, with Sim(Ref) of 0.555 versus 0.364 and Sim(Tgt) of 0.461 versus 0.316, and composition, with Count of 0.869 versus 0.828 and Coverage of 0.960 versus 0.813. It also adds ID tokens, Representation Forcing, and the LG-ID Loss, however, so this comparison does not isolate the contribution of the corpus.

P4 adds an ID token to P1, and P5 adds Representation Forcing on top of P4. The ID token alone yields small but consistent gains, the largest of which is Count, from 0.771 to 0.827. Representation Forcing then raises Sim(Ref) from 0.351 to 0.364 and Sim(Tgt) from 0.313 to 0.328. In the three examples shown in Figure 8, each supervised identity prediction preferentially attends to its corresponding ID token. The end-to-end gains are modest, so we interpret this evidence as consistent with improved identity addressability rather than as a large direct source of identity gain, and the attention analysis does not establish the effect relative to a model without Representation Forcing, nor equally strong routing from generated face patches to the ID tokens.

C Experimental Details

C.1 Training Sequence and Data Construction

The causal training sequence is organized as follows: **1)** reference images and the prompt; **2)** reference-identity selection and ID loading; **3)** face layout and identity binding, followed by body-region and pose planning; **4)** rendering the plan into a visual layout condition; **5)** a summary and detailed recaption; **6)** target identity representation prediction; and **7)** target image generation. Stage 5 restates the request after the plan has been committed, giving the generator a next-token-supervised language description that is already consistent with the rendered canvas instead of only the short user prompt. This order is the one used in training and inference; Figure 2 arranges some stages differently for visual layout only, and the order stated here takes precedence.

Each group-image training sample contains a target image, reference images, and reference-to-target identity correspondences, with short captions as user prompts and detailed captions as recaptions. Face detection and ArcFace provide reference and target identity embeddings, while person detection and pose estimation by YOLOv11 (Ultralytics, 2024) provide person regions and keypoints. Together with the captions and correspondences, these signals supervise the structured reasoning stages and let the renderer produce the layout condition. We additionally use text-to-layout examples to supervise layout reasoning without a target image, strengthening planning independently of full image-generation samples.

C.2 Identity Loading and Representation Forcing

The prompt may require only a subset of the available reference people and may also describe people without a reference image, so the model first identifies the participating identities. Selection and binding are ordered rather than mutually dependent: the model emits the list of selected identities and loads one ID token per selection, and the binding of each token to a concrete planned person follows in the layout stage. This identity–layout binding gives each loaded identity a spatial role rather than treating the references as an unordered collection. The ID embeddings complement the low-level reference-image features with high-level identity-specific information, analogous to the complementary roles of ViT (Dosovitskiy et al., 2020) and VAE embeddings at different feature levels: for each selected person, a 512-dimensional ArcFace embedding is mapped to the model hidden dimension by a lightweight MLP and injected at the corresponding ID-token position.

Loading ID tokens does not by itself guarantee that image generation will use them. Each person is represented by a single continuous token, projected from one 512-dimensional vector, inside an interleaved sequence of 10–20K tokens, so generation-side attention over the injected identity tokens can be diffuse and identity-agnostic. For every selected reference identity with a target

correspondence, Representation Forcing therefore places a representation token before the target image. The backbone computes its hidden state from the preceding prompt, visual references, ID tokens, and layout reasoning, and the output projector g_{out} maps it to the ArcFace space, where it is compared with the ArcFace embedding of the corresponding person in the target image. The resulting hidden states remain causally available to the later image tokens, forming an identity scaffold for generation.

C.3 Training Configuration

WithEveryone uses a mixture-of-transformers backbone with 60B parameters on the understanding side and 60B parameters on the generation side, initialized from an in-house base model. Training uses 400K in-house group-image samples; Figure 3 reports the identity-occurrence and reference-count statistics of this corpus. We optimize with Muon (Liu et al., 2025a) at a learning rate of 1×10^{-5} on the generation side and 3×10^{-6} on the understanding side, with a packed sequence length of 72K tokens. Training runs on 128 H20 GPUs, and all reported models are taken from 1600 iterations. The objective weights are $\lambda_{\text{FM}} = 1.0$, $\lambda_{\text{RF}} = 1.0$, and $\lambda_{\text{ID}} = 0.5$.

C.4 Shared Evaluation Protocol

Existing multi-identity benchmarks focus on a small number of people and therefore do not measure how identity preservation scales to larger groups, which is why we construct the five-to-ten-person benchmark of Section 4.1. All models are scored on the same 210 benchmark examples with the same detection, matching, and aggregation code; no example is excluded for any model. Each evaluation example contains one generated image, 5–10 reference identities, a target group image, and reference-to-target identity correspondences. Face detections and embeddings are computed independently for the generated, reference, and target images. For models with Layout CoT, we additionally parse face boxes, body boxes, pose keypoints, and identity–layout bindings from the generated plan. Body-level evaluation uses YOLO11x-pose with a confidence threshold of 0.25 and an input size of 640; if body detection is unavailable, only face-level terms are retained.

C.5 Identity Metrics

The main comparison follows the WithAnyone protocol and averages results from ArcFace, FaceNet, and AdaFace. For each backbone, generated faces are matched to reference or target faces by maximum-similarity Hungarian assignment. The resulting means define Sim(Ref) and Sim(Tgt), respectively. Copy-Paste, introduced and validated against human judgment by WithAnyone (Xu et al., 2025), measures the scale of copy-paste artifacts. It compares each matched generated–target–reference triplet in angular embedding space:

$$\text{Copy-Paste} = \frac{\theta_{gt} - \theta_{gr}}{\theta_{tr}}, \tag{5}$$

where $\theta_{ab} = \arccos(\cos(\mathbf{e}_a, \mathbf{e}_b))$. A larger value indicates that the generated face is closer to the reference than to its target-context appearance. The ablation Sim(Ref) and Sim(Tgt) metrics use the same generated-to-reference and generated-to-target semantics with a single ArcFace encoder.

C.6 Image and Text Similarity Metrics

CLIP-I and DINO-I are the cosine similarities between the generated and the target image in the CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) image-embedding spaces, and CLIP-T is the cosine similarity between the generated image and the user prompt in the joint CLIP space. These three metrics describe overall scene and prompt agreement and are insensitive to which identity appears where, so we read them alongside the identity metrics above rather than on their own.

C.7 Layout Score

Layout Score measures the overall validity of the final generated image as a weighted aggregate of seven sub-scores in $[0, 1]$. If a component cannot be computed, it is omitted and the remaining weights are renormalized.

Count. Let N be the expected number of people in the target and M the number of detected generated faces. We compute

$$S_{\text{count}} = 1 - \min\left(1, \frac{|M - N|}{\max(1, N)}\right), \tag{6}$$

so both missing and extra people are penalized relative to the target group size.

Coverage, uniqueness, and distinctness. We construct the cosine-similarity matrix between reference and generated faces. Maximum-similarity Hungarian assignment determines one-to-one matches, and a reference is covered when its matched similarity is at least 0.20; S_{cover} is the fraction of covered references. For uniqueness, a reference is marked as over-claimed if more than one generated face has similarity at least 0.20 to it, and S_{uniq} is one minus the over-claimed-reference rate. Distinctness captures the opposite collision: each reference independently selects its most similar generated face, and S_{distinct} is one minus the fraction of references whose best face is also selected by another identity. The Dup column of Table 1 reports this distinctness collision rate, $1 - S_{\text{distinct}}$.

Sensitivity to the matching threshold. The 0.20 similarity threshold is permissive by design, so that a reference counts as covered whenever the generated person is recognizably the intended one. Coverage is nevertheless insensitive to it: for P7, raising the threshold to 0.25, 0.30, and 0.35 lowers Coverage by 0.010, 0.017, and 0.027, and lowering it to 0.15 raises Coverage by 0.012. Only 3.6% of references have a matched similarity in $[0.15, 0.30)$, against a median matched similarity of 0.554, so few decisions sit near the boundary.

Anonymous leakage and spatial validity. Generated faces that are not valid Hungarian matches are treated as anonymous. An anonymous face is counted as identity leakage if its similarity to any reference is at least 0.25; S_{noleak} is one minus the leaked-anonymous-face rate. Spatial validity is computed from pairwise overlaps. Face IoU above 0.30 and body IoU above 0.50 are treated as severe overlaps. When both detections are available, S_{spatial} is one minus the mean of the face- and body-overlap rates; otherwise, only the face-overlap rate is used.

Plan adherence. We match generated face boxes to the available layout boxes and average the matched IoUs. Ground-truth boxes are used when a ground-truth layout is available; otherwise, boxes parsed from the model's CoT plan are used. This mean IoU defines S_{plan}.

C.8 Relative Layout Score and Plan IoU

The Relative Layout Score (RLS), adopted from "R-spatial" in T2I-R1 (Jiang et al., 2026), evaluates the relative configuration of the model's textual plan against the ground-truth layout, without using the generated image. It combines relative count, position, and size, all in $[0, 1]$ and higher-is-better.

Person matching. Planned and ground-truth people with the same reference identity are matched directly. Remaining people are represented by face centers, or body centers when faces are unavailable. The centers in each layout are centered and normalized by their root-mean-square spread before Hungarian geometric matching, preventing global translation or scale from dominating correspondence.

Relative count. Face and body counts are scored independently with the same normalized cardinality rule as S_{count} above, using planned and ground-truth counts. The available face- and body-count scores are then averaged into $S^{\text{rel}}_{\text{count}}$.

Relative position. For every pair of matched people, we check whether their left–right and top–bottom ordering agrees between the plan and ground truth, giving the two order-consistency terms $S_{\text{order-}x}$ and $S_{\text{order-}y}$. We additionally align planned centers to ground-truth centers with a non-reflective Umeyama similarity transform and convert the normalized residual distances into an OKS-style soft alignment score S_{align}. A fitted-scale penalty prevents a collapsed layout from receiving a high score merely because Procrustes alignment can enlarge it. The order-consistency and aligned-center scores jointly define the position term,

$$S_{\text{order}} = 0.6\, S_{\text{order-}x} + 0.4\, S_{\text{order-}y}, \qquad S^{\text{rel}}_{\text{pos}} = 0.6\, S_{\text{order}} + 0.4\, S_{\text{align}}, \tag{7}$$

where the horizontal ordering receives the larger of the two order weights.

Relative and absolute size. For each matched face or body box, its area is divided by the mean matched-box area in the same layout. The clipped absolute log-ratio between planned and ground-truth normalized areas measures relative-size agreement. We also compare the mean absolute face area between the two layouts to retain a penalty when every planned person is uniformly too large or too small. The available face, body, and absolute-size scores are averaged into $S^{\mathrm{rel}}_{\mathrm{size}}$.

Aggregation. The three terms are combined as

$$\mathrm{RLS} = 0.20\, S^{\mathrm{rel}}_{\mathrm{count}} + 0.45\, S^{\mathrm{rel}}_{\mathrm{pos}} + 0.35\, S^{\mathrm{rel}}_{\mathrm{size}}. \tag{8}$$

Because the position term is measured after similarity alignment and the size term is normalized within each layout, RLS is insensitive to a global translation, rotation, or scaling of the whole plan and instead scores the arrangement of people relative to one another. It does not penalize the person overlap that is normal in a group photograph, and it is not an absolute bounding-box IoU.

Plan IoU instead evaluates the generation side. Planned face boxes are matched to detected faces in the generated image, and the mean matched IoU is reported. When ground-truth layout is directly provided as the generation condition, the same computation uses ground-truth boxes. Thus, RLS measures whether the understanding side proposes an appropriate relative layout, whereas Plan IoU measures whether image generation executes the supplied plan, and Layout Score measures the validity of the final image, of which plan adherence is only one weighted sub-score.

D Extended Background

Generative backbones. Text-to-image generation was established by denoising diffusion models (Ho et al., 2020) built on convolutional U-Net denoisers (Ronneberger et al., 2015) and moved into a latent space to make high-resolution synthesis tractable (Rombach et al., 2022). Later systems replaced the convolutional denoiser with transformer architectures (Vaswani et al., 2017; Peebles & Xie, 2023) and the discrete noise schedule with flow matching (Lipman et al., 2022), the recipe followed by current open-weight models (Labs, 2024; 2025b; Batifol et al., 2025). Transformer backbones are permutation-equivariant over their token sequence up to positional encoding (Xu et al., 2024), which is what lets heterogeneous conditions such as reference images, ID tokens, and layout tokens be interleaved in a single sequence. Our generation side inherits this design, but its conditioning signals are produced by the understanding side of the same model rather than by an external encoder. Contrastively pre-trained image–text encoders (Radford et al., 2021; Zhai et al., 2023) remain the standard source of semantic image features and of the CLIP-based metrics reported in Section 4.

Reference-conditioned and identity-preserving generation. Before unified models, reference conditioning was usually attached to a frozen generator: adapters inject reference features through additional cross-attention (Ye et al., 2023), and side networks inject spatial conditions (Zhang et al., 2023). Face-specific variants specialize this recipe to identity (Papantoniou et al., 2024; Peng et al., 2024), and attention routing extends it from one subject to several (Wang et al., 2024a; Guo et al., 2025). A parallel line folds these tasks into a single model that generates and edits from interleaved inputs (Xiao et al., 2024; Liu et al., 2025b; Wu et al., 2025d;c). Our setting differs in scale rather than in conditioning mechanism: with five to ten identities in one image, the limiting factor becomes how each identity is supervised and addressed, not how reference features enter the network.

Planning and representation objectives. Layout planning has also been delegated to a language model that converts the prompt into a layout for a separate diffusion renderer (Lian et al., 2023), whereas we keep planning inside the same generative context so that the plan and the image share one state (Section 3.2). Supervising an intermediate prediction against a target representation is a long-standing idea in representation learning (Oord et al., 2018), which Section 3 specializes to one supervised prediction per reference identity rather than a single global representation.

E LIMITATIONS, FUTURE WORK, AND RESPONSIBLE USE

Layout is hard to evaluate. We evaluate layout from two perspectives: whether the faces are correctly planned, and whether the plan follows the relative location relationships of the ground truth, for example whether the person on the left remains on the left. When the prompt is underspecified, however, which is common in real-world use, many layouts may be equally valid and layout quality becomes difficult to measure against a single reference. We leave this to future work.

Evaluation scope. Our conclusions are drawn from 210 examples on a single benchmark, and the groups with more references are the smallest, so per-group numbers should be read as a trend rather than as precise estimates. All identity measures inherit the behavior of the underlying face detectors and recognizers, whose accuracy is known to vary across demographic groups, and the compared systems differ in their maximum output resolution and in whether they can plan a layout on their own.

Data and responsible use. Training uses group images collected from publicly accessible sources solely for non-commercial research, as described in the Ethics statement. As with other identity-preserving generators, including the proprietary systems we compare against, identity-conditioned generation can be misused to produce images of real people without their consent, to place someone in a scene they never took part in, or to impersonate them. These risks are common to the task rather than specific to our method; deployments should require consent from the referenced people and pair generation with provenance signaling.