COGCANVAS: A BENCHMARK FOR EVALUATING MULTI-SUBJECT REFERENCE-BASED IMAGE GENERATION

Anonymous authors
Paper under double-blind review



Figure 1: **CogCanvas composing scenes from joint identity, object, and background references.** Each prompt provides $N \in \{2, \ldots, 5\}$ identity references, per-person object/fashion references, optional inter-person interaction annotations, and a background scene reference; the model must produce a single photorealistic image that preserves every identity, binds the correct object to each person, follows the prescribed spatial layout, and realizes the stated interactions. This three-reference setting is harder than prior multi-subject benchmarks: existing methods, tuned for human-only or human–object personalization, consistently ignore the background reference, and attribute–person binding collapses as N grows. **CogCanvas** exposes this gap through a six-axis protocol, *ID-Sim*, *DINO-Sim*, *CLIP-T*, *BG-Sim*, *Aesth.*, and *Attr-VQA*, and a training-free reference pipeline that accepts all three reference modalities via sequential inpainting.

ABSTRACT

Multi-subject reference-based image generation requires jointly preserving multiple human identities, binding per-person objects and fashion items, and respecting a specified background scene, a regime where current diffusion models remain brittle. Existing benchmarks evaluate only one axis at a time and none jointly captures multi-identity composition with human-object interaction, background grounding, and spatial plausibility. We introduce **CogCanvas**, a benchmark of 1,952 curated reference images spanning 100 celebrity identities, 115 distinctive objects and fashion items, and 29 real-world background scenes including Vietnamese landmarks, from which we construct 1,361 compositional prompts covering 2–5 person group sizes. The curation pipeline combines DINOv2-based deduplication, automated filtering, human review, and structured interaction and position annotations that make individual relations checkable. **CogCanvas** supports three tasks, reference-based multi-human-object generation (primary), text-to-image compositional generation, and reference retrieval, under a unified six-axis evaluation protocol. We introduce two metrics tailored to the multi-reference setting: *BG-Sim*, which scores background fidelity on SAM 3-masked regions via DINOv3 feature similarity, and *Attr-VQA*, which uses a multimodal LLM to verify per-subject attribute binding and inter-person interactions against the structured graphs. Benchmarking five existing methods and a training-free sequential-inpainting reference pipeline shows that object and fashion fidelity drops sharply as group size grows from two to five. A 46-image human study finds moderate agreement between Attr-VQA and human judgments (Spearman $\rho = 0.645$).

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.

1 INTRODUCTION

The ability to generate images containing multiple identity-specific people from a single text prompt represents a frontier challenge for modern Text-to-Image (T2I) generation systems. Recent diffusion-based models such as FLUX Black Forest Labs (2024), Stable Diffusion Rombach et al. (2022), and SeedDream Gao et al. (2025) have demonstrated remarkable single-subject personalization capabilities. However, extending these capabilities to scenes with 2–5 distinct, identity-specific individuals introduces a cascade of compounding difficulties: the model must simultaneously preserve the unique visual identity of each person, correctly bind outfit attributes and objects to the specified individuals, place them in accurate spatial configurations (*e.g.*, "A stands to the left of B"), and realize meaningful physical interactions (*e.g.*, "A and B shake hands").

A fundamental bottleneck to progress in this area is the absence of a benchmark that captures these challenges in an integrated, systematic way. Existing benchmarks either focus on single-subject fidelity Ruiz et al. (2023), object-level composition Huang et al. (2023), or evaluate spatial relationships only as binary checks between pairs of objects (Table 1). None of them simultaneously evaluate multi-hop spatial reasoning, human-human interaction correctness, and fine-grained attribute binding in multi-person scenes. MultiHuman-Testbench Borse et al. (2025) addresses the multi-human generation setting but does not evaluate attribute-object binding, and its prompts describe simple group actions rather than complex spatial arrangements and interactions.

To fill this gap, we introduce **CogCanvas**, a comprehensive benchmark for *Multi-Subject Reference-Based Image Generation*. Given a set of reference images for each identity, the task requires the model to compose a scene that jointly preserves all specified identities, binds the correct attributes and objects to each person, and realizes the described spatial arrangements and interactions.

CogCanvas is built on three organizing principles. First, *diversity*: references span 100 celebrity identities, 80 unique objects, 35 fashion items, and 29 distinctive background scenes (*e.g.*, real-world famous landmarks such as Ha Long Bay, Dragon Bridge, and Fansipan), providing individually identifiable contexts in which background-reference fidelity can be checked. Second, *multi-reference complexity*: each prompt involves 2–5 persons, each paired with specific object/fashion references, simultaneously testing identity preservation and attribute binding. Third, *logic-focused evaluation*: ground-truth interaction and position graphs enable fine-grained verification of interaction fidelity and spatial layout accuracy via structured visual question-answering (VQA) queries, rather than relying solely on perceptual quality.

Our main contributions are as follows:

- **Benchmark and curation pipeline. CogCanvas** provides 1,952 curated reference images and 1,361 compositional prompts (2–5 persons) jointly covering identity, object/fashion, and background references, paired with a multi-modality curation workflow that automatically derives structured interaction and position graphs as ground-truth supervision.
- **Tasks and metrics.** We define three tasks, including reference-based multi-subject generation, text-to-image compositional generation, and reference retrieval, and introduce two new metrics: BG-Sim (SAM-3 + DINOv3 background fidelity) and Attr-VQA (MLLM verification of per-subject attribute binding and interaction graphs).
- **Baseline and benchmarking.** A training-free sequential-inpainting pipeline assembled from existing components supplies a runnable reference point for all six axes. Five existing methods exhibit sharp declines in identity and object/fashion fidelity as group size increases; a human study measures how well Attr-VQA tracks human judgments.

2 RELATED WORK

Image Generation & Reference-Based Methods. Diffusion-based models have driven remarkable progress in text-to-image synthesis, with the Stable Diffusion series Rombach et al. (2022); Podell et al. (2023) establishing the foundation for personalization research, and FLUX Black Forest Labs (2024) advancing further via a Multimodal Diffusion Transformer (MM-DiT) Esser et al. (2024) that jointly attends over text and image tokens. Flux-Kontext Labs et al. (2025) extends this framework to reference-conditioned generation without fine-tuning. Building on such backbones, recent methods condition generation on visual references to preserve subject identity: OmniGen2 Wu et al. (2025a) and UNO Wu et al. (2025b) handle multiple references through unified training and shared visual

context, while DreamO Wu et al. (2025c) and XVerse Chen et al. (2025) use dedicated subject embeddings, and MOSAIC She et al. (2026) improves attribute binding via region-aware attention at inference time. Despite these advances, all methods degrade significantly as the number of subjects increases-a failure mode our benchmark systematically quantifies.

Text-to-Image Benchmarks. T2I benchmarks are categorized by the granularity of their annotations. HPD Wu et al. (2023) relies on coarse human preference scores, while GenEval Ghosh et al. (2023) and T2I-CompBench Huang et al. (2023) introduce task-specific accuracy metrics for compositional attributes. EvalMuse-40K Han et al. (2024) evaluates aesthetics and text alignment at scale. However, these benchmarks treat spatial relationships as simple binary checks between two objects and do not evaluate multi-person identity-specific generation. DreamBench Ruiz et al. (2023) and DreamBench++ Peng et al. (2025) focus on single-subject personalization and do not address multi-subject scenes. XVerseBench Chen et al. (2025) extends to multi-subject settings (up to 3 subjects) but does not evaluate interaction correctness or complex spatial configurations. **CogCanvas** advances the field by introducing ordered spatial permutations ($N!$) of group layouts and structured interaction graphs, requiring models to demonstrate understanding of both scene composition and identity-specific attribute binding.

Multi-Human Generation Benchmarks. MultiHuman-Testbench Borse et al. (2025) is the closest existing work to ours. It provides 1,800 samples with 5,550 reference face images and evaluates face count, Hungarian ID similarity, prompt alignment, and action detection. However, its prompts describe simple group actions without fine-grained spatial arrangements, and it does not evaluate attribute binding (outfits, objects) or interaction graphs. Our benchmark specifically targets the harder setting of compositional generation, where each identity must carry specific object and fashion references, and spatial arrangements must be verified at the permutation level.

3 CogCanvas Dataset

3.1 Overview

CogCanvas evaluates *Multi-Subject Reference-Based Image Generation*: given a set of curated reference images for each specified identity, models must compose a scene that preserves all identities, assigns the correct object and fashion attributes to each person, places subjects in accurate spatial configurations, and realizes the described interactions. The reference database spans multiple categories, including human identities, general objects, fashion items, and background scenes, with controlled variation in pose, scale, and appearance.

The benchmark is designed around three core principles: (i) **diversity**: references span multiple categories with controlled variation in identity appearance, object type, and background scene; (ii) **multi-reference complexity**: prompts require models to jointly handle 2–5 distinct identities, each paired with specific object/fashion references; and (iii) **logic-focused evaluation**: ground-truth interaction and position graphs enable formal verification of interaction fidelity and spatial layout accuracy rather than relying solely on perceptual metrics.

3.2 Reference Curation

We partition reference images into three disjoint groups and curate each independently.

Person. We select 100 celebrity identities and their corresponding high-resolution portraits from CelebA-HQ Karras et al. (2018), a widely used dataset for training generative models of human faces. Each identity is represented by multiple images capturing diverse poses and expressions. Annotators audit the joint distribution of gender, age group, and ethnicity and replace candidates to improve balance within this source. This best-effort audit does not make the pool representative of the wider population (Appendix B).

Object & Fashion. We curate 80 general objects and 35 fashion items (115 unique items), using DINOv2 (Oquab et al., 2024) features to deduplicate web-collected images. We choose distinctive items such as "Ao Dai Hue" rather than generic categories such as "blue sweater", so an arbitrary category exemplar cannot satisfy visual matching. A two-stage filter applies aesthetic and relevance checks. The pool then undergoes a *manual review* for distinctiveness and image quality; candidates with blur, glare, or heavy occlusion are replaced.

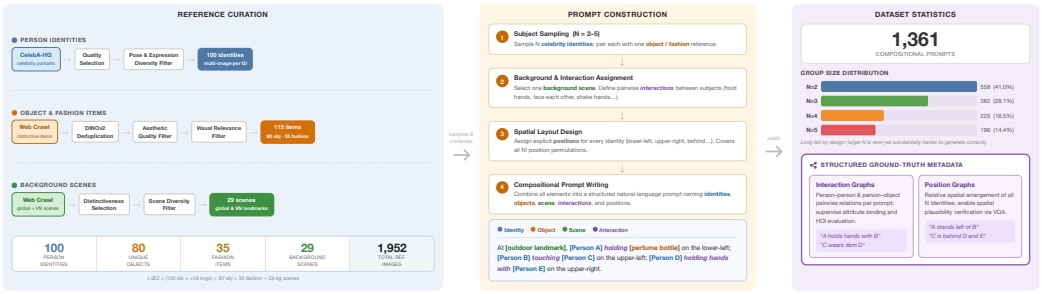


Figure 2: **CogCanvas data curation pipeline.** *Left:* Three disjoint reference pools are curated independently: person identities from CelebA-HQ, object/fashion items via web crawling with DINOv2 deduplication and two-stage filtering, and background scenes selected for global diversity. *Center:* Compositional prompts are constructed by sampling $N \in \{2, \ldots, 5\}$ identities, assigning per-person object/fashion references, selecting a background scene, defining pairwise interactions, and specifying explicit spatial layouts covering all $N!$ position permutations. *Right:* The resulting 1,361 prompts follow a long-tail group-size distribution and are annotated with structured interaction and position graphs that serve as ground-truth supervision for attribute-binding and spatial-plausibility evaluation.

Background. We collect 29 distinctive background scenes depicting real-world environments, including famous landmarks (*e.g.*, Ha Long Bay, Dragon Bridge, Fansipan, Ba Na Hills) and natural landscapes. Each scene is individually identifiable, so background matching tests the supplied reference rather than a generic scene category. Distinctive interiors could satisfy this criterion, but generic rooms are outside the current pool.

3.3 Prompt Construction

We design the prompt set to stress-test *human–human interaction*, *multi-reference composition*, and *spatial reasoning*. Each prompt describes a scene of 2–5 persons drawn from the celebrity identity pool, paired with object or fashion references and grounded in a specified background scene, so that generation models must jointly handle identity preservation and attribute binding within the scene.

Attribute extraction. A multimodal LLM (MLLM) extracts structured metadata for every reference asset: *age*, *gender*, and *ethnicity* per identity, and *size category* (small handheld $\rightarrow$ large wearable) together with *afforded interactions* (*e.g.*, hold, wear, stand-beside) per object or fashion item. The typed pool lets the sampler condition on attributes rather than raw IDs.

Compositional schema. We formalize each prompt as the tuple $(B, \{H_i\}_{i=1}^{N}, \{O_j\}_{j=1}^{M}, \{R_j\}_{j=1}^{M})$ with $M \leq N$, where B is a background reference, $\{H_i\}$ are N identity references, $\{O_j\}$ are M object or fashion references each bound to a single person, and $\{R_j\}$ are the corresponding human–object interaction specifications. This schema reduces every prompt to a verifiable set of bindings rather than free-form text, and is the structure that the MLLM-based pipeline below operates over.

Diversity-controlled sampling and synthesis. We sample joint tuples uniformly at random across two controllable axes, such as *object size* (small to large) and *interaction strength* (close-contact "holding" to loose "standing beside"), so the prompt set spans the full spectrum of compositional difficulty rather than concentrating on easy modes. Conditioned on each tuple, the MLLM writes a natural-language prompt that grounds the identities in the background and binds each person to her assigned object through a reasoned interaction phrase.

Verification. Every synthesized prompt then passes an MLLM verification stage that checks (i) *interaction plausibility*-is the human–object interaction consistent with the object's affordances?; (ii) *layout feasibility*-can N persons be co-placed without unrealistic overlap in the specified background?; and (iii) *scene compatibility*-does the background admit the specified group activity? Prompts failing any check are regenerated until they pass; 1.2% of candidates require a human override. Together with manual reference screening and the demographic audit, curation involved roughly 60 person-hours across two annotators (Appendix D). Human review occurs during dataset construction, not during model inference.

The resulting set comprises 1,361 prompts distributed across group sizes: 2-person (558, 41.0%), 3-person (382, 28.1%), 4-person (225, 16.5%), and 5-person (196, 14.4%). This long-tail distribution

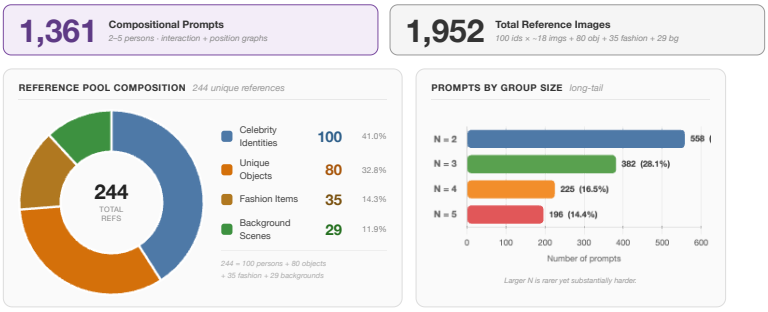


Figure 3: **CogCanvas dataset statistics.** *Top:* Total compositional prompts (1,361) and reference images (1,952). *Left:* Pie chart showing the composition of the 244 unique references by category: 100 celebrity identities (41.0%), 80 unique objects (32.8%), 35 fashion items (14.3%), and 29 background scenes (11.9%). *2:* Long-tail prompt distribution across group sizes N = 2–5.

reflects real-world complexity, where larger compositions are naturally rarer yet substantially more challenging to generate correctly.

3.4 Metadata Generation

To enable fine-grained benchmarking beyond perceptual quality, Gemma 4 generates prompt text and structured interaction and position metadata from the sampled reference tuples. These synthetic fields are verified before release; they are not independent human labels. The resulting annotations are:

Reference Metadata. For every reference asset we release the structured attribute fields produced by the MLLM extractor described in Section 3.3: *age*, *gender*, and *ethnicity* for identity references; *size category* (small handheld → large wearable) together with *afforded interactions* (*e.g.*, hold, wear, stand-beside) for object and fashion references; and *scene type* together with *landmark name* for background references. This per-reference metadata is the foundation on which the prompt-level graphs below are built, supports the reference retrieval subtask (Section 3.5), and lets evaluators slice benchmark results along controllable axes such as object size or demographic group.

Interaction Graphs. For each prompt, we construct a graph that encodes pairwise relationships between persons and between persons and objects (*e.g.*, "A holds hands with B", "C wears item D", "A and B face each other"). These graphs provide ground-truth supervision for evaluating interaction fidelity in generated images.

Position Graphs. For each prompt, we capture the relative spatial arrangement among all identities in the scene (*e.g.*, "A stands to the left of B", "C is positioned behind D and E"). Together with interaction edges, these annotations turn each specified relation into a separate subject-linked binary question for Attr-VQA (Section 4), allowing errors to be localized beyond a prompt-level CLIP-T score.

3.5 Dataset Statistics

CogCanvas comprises 1,361 compositional prompts and 1,952 reference images spanning 100 celebrity identities, 80 objects, 35 fashion items, and 29 background scenes; modality breakdown and the 2-to-5-person group-size distribution are summarized in Figure 3.

A distinctive property of **CogCanvas** is that *each identity is paired with a pool of reference images* captured under varied lighting, pose, and apparel, so that different references can be attribute-matched to a given prompt-a subject described as "wearing a red jacket" should be grounded to the reference that shows that outfit, not an arbitrary one. This design enables a *reference retrieval subtask*: select the highest-fidelity reference for each named slot given the prompt, serving as a diagnostic of fine-grained text-to-reference grounding before generation.

Table 1 shows the benchmark's emphasis on joint coverage rather than prompt count: the 1,361 prompts are a curated sample of compositions formed from 100 identities, 115 object/fashion items, and 29 scenes. Each prompt carries several separately verifiable identity, binding, position, and interaction assertions.

Table 1: Comparison of text-to-image and multi-subject generation benchmarks. **#Prompts**: number of evaluation prompts. **Human**: number of human identity references (faces). **Object**: number of object/animal references. **HOI**: explicit human–object interaction evaluation. **BG**: background scene reference provided. **Spatial**: spatial plausibility evaluation between subjects and/or background. **CogCanvas** is the only benchmark covering all five dimensions simultaneously.

Benchmark	Venue	#Prompts	Human Refs	Object Refs	HOI	BG	Spatial
T2I-CompBench	NeurIPS'23	6,000	0	0	No	No	Partial
DreamBench	CVPR'23	750	0	30	No	No	No
DreamBench++	ICLR'25	1,350	20	110	No	No	No
MultiHuman-TB	NeurIPS'25	1,800	5,550	0	Partial	No	No
XVerseBench	NeurIPS'25	300	20	119	Partial	No	No
CogCanvas	**Ours**	**1,361**	**100**	**115**	**Yes**	**Yes**	**Yes**

Counts: DreamBench = 30 subjects (9 animals + 21 objects) × 25 prompts; DreamBench++ = 150 subjects (20 humans + 45 animals + 65 objects) × 9 prompts; XVerseBench = 139 subjects (20 humans + 45 animals + 74 objects) across 300 prompts.

4 Tasks and Evaluation Metrics

The primary benchmarking task is **reference-based multi-human-object generation**: given a compositional text prompt and per-subject reference images, generate a scene that faithfully reproduces all specified identities, interactions, and background context. We evaluate across six complementary dimensions below. The dataset structure additionally supports two auxiliary tasks, including text-to-image compositional generation and reference retrieval, described at the end of this section; we release their evaluation protocols but do not run a full model benchmark on them.

Face ID Similarity (ID-Sim). We compute cosine similarity between ArcFace Deng et al. (2019) embeddings of reference and generated faces, averaging over all persons in the prompt. This metric directly measures identity preservation: whether each generated face matches the specified reference identity. In multi-subject scenes, we use Hungarian matching Kuhn (1955) to align detected generated faces to reference identities before computing similarity, preventing inflated scores from incorrect identity pairings.

DINOv2 Similarity (DINO-Sim). We compute cosine similarity between DINOv2 (Oquab et al., 2024) features of reference objects/fashion items and their generated counterparts. Unlike face embeddings, DINOv2 captures holistic semantic similarity for non-face entities such as specific object models and clothing items, assessing whether the generated image faithfully reproduces their key visual attributes.

CLIP-Text Score (CLIP-T). We measure CLIP Radford et al. (2021) similarity between the input text prompt and the generated image as a measure of overall text-image semantic alignment. This metric captures whether the generated scene broadly matches the described content, but is insufficient alone as it cannot distinguish correct identity assignments from swapped ones.

Aesthetic Score (Aesth.). We apply the LAION aesthetic predictor Schuhmann (2022) to measure the perceptual quality of generated images, independent of content accuracy. This metric evaluates whether the generation method produces visually natural and pleasing images, which is necessary but not sufficient for benchmark success.

Background Consistency Score (BG-Sim). BG-Sim measures background fidelity on foreground-masked regions. We use SAM 3 Carion et al. (2025) to mask all foreground subjects in both the generated image and the reference background, extract DINOv3 Siméoni et al. (2025) features from the remaining regions, and report their cosine similarity averaged over prompts. Unlike CLIP-T, whose global alignment is easily dominated by salient subjects, BG-Sim directly penalizes hallucinated or inconsistent backgrounds, *e.g.*, an indoor scene rendered outdoors, or a generic studio backdrop replacing a specified urban environment.

Attribute VQA Score (Attr-VQA). Attr-VQA uses a multimodal LLM (Gemma4 Manik & Wang (2026)) to verify per-subject semantic fidelity via binary Yes/No questions automatically generated from the structured metadata (Section 3.4). We probe two complementary aspects: *attribute binding*,

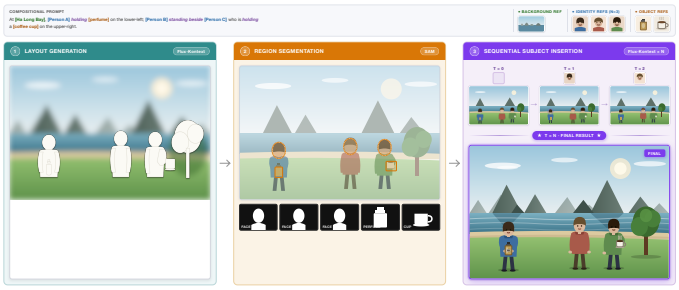


Figure 4: **Training-free sequential-inpainting reference pipeline.** Given a compositional prompt with background, identity, and object/fashion references, the pipeline generates a scene layout with subject placeholders, segments face/object regions, and sequentially inpaints each region conditioned on its corresponding reference to reduce cross-subject feature leakage.

which checks that visual attributes are bound to the correct identity rather than leaking across subjects (*e.g.*, "Does the person at [position] wear a [color] outfit?"), and *interaction consistency*, which checks that generated images realize the ground-truth interaction and position graphs (*e.g.*, "Is person A holding hands with person B?"). The final score is the mean accuracy across all questions, entities, and prompts. The question bank tests attributes, positions, and interactions; it does not test lighting direction, cast shadows, or physical coherence, for which the prompt graphs provide no ground truth. Attr-VQA is the only one of the six reported axes that uses an MLLM.

Additional supported tasks. **CogCanvas** also supports two auxiliary tasks that we release as evaluation protocols without a full benchmark. *Text-to-image generation*: prompts can be used without reference images to test compositional scene generation from text alone; evaluation uses human count accuracy (person-detector count vs. ground-truth count) and Attr-VQA restricted to attribute-binding queries. *Reference retrieval*: because each identity ships with a pool of attribute-varied reference images, the task of selecting the best-matching reference for each identity slot in a prompt can be evaluated via Precision@k and Recall@k ($k \in \{1, 3, 5\}$), where the ground truth is the annotated attribute-matched image per slot.

5 Benchmarking

5.1 Proposed Baseline

We use a *training-free* sequential-inpainting pipeline, assembled from existing components, as a reproducible reference point that can accept identity, object/fashion, and background inputs (Figure 4). It is a benchmark control rather than a new generation method. Conditioning each pass on one subject reference reduces opportunities for cross-subject leakage, while inference latency grows approximately linearly with the number of subjects N. This comparison is diagnostic, not a controlled proof that leakage alone causes the observed declines.

The pipeline proceeds in three stages. **Stage 1 - Layout Generation**: Given the compositional prompt and a background scene reference image, Flux 2.0 Klein Black Forest Labs (2026) generates a *base image* containing N anonymous face placeholders whose positions and spatial relations conform to the spatial constraints described in the prompt. No identity references are supplied at this stage; the model is asked only to establish a plausible scene layout over the provided background. **Stage 2 - Region Segmentation**: SAM Kirillov et al. (2023) is applied to the base image to produce binary masks for each face and associated object/fashion region, yielding one inpainting target per subject slot. **Stage 3 - Sequential Subject Insertion**: For each subject $i = 1, \ldots, N$, Flux 2.0 Klein performs a masked inpainting pass conditioned on subject i's reference image and SAM mask, replacing its placeholder; processing subjects one at a time isolates each conditioning step and prevents cross-subject feature leakage.

5.2 Experimental Setup

Benchmark Methods. We compare the proposed baseline with representative multi-subject generation methods: OmniGen2 Wu et al. (2025a), UNO Wu et al. (2025b), DreamO Wu et al. (2025c), XVerse Chen et al. (2025), and MOSAIC She et al. (2026).

Table 2: Quantitative comparison on **CogCanvas** by group size N. Seq. ref. is the training-free sequential reference pipeline. Best results are in **bold**, second-best underlined. ↑: higher is better. BG-Sim is reported only where the method accepts a background reference.

N	Method	ID-Sim ↑	DINO-Sim ↑	CLIP-T ↑	BG-Sim ↑	Aesth. ↑	Attr-VQA ↑
2	OmniGen2	13.07	**33.16**	62.23	–	51.63	33.04
	UNO	23.23	28.03	62.11	–	55.24	35.12
	DreamO	3.07	5.98	58.79	–	44.42	14.45
	XVerse	42.33	32.96	61.60	–	39.73	7.80
	MOSAIC	9.93	4.77	57.29	–	44.01	11.22
	Seq. ref.	**45.28**	32.02	**64.07**	**0.732**	**55.33**	**53.41**
3	OmniGen2	0.62	1.74	57.57	–	37.11	3.68
	UNO	19.60	17.35	61.02	–	**56.01**	14.65
	DreamO	0.39	1.20	57.47	–	41.48	2.99
	XVerse	27.21	15.62	60.08	–	38.46	4.03
	MOSAIC	11.45	5.87	57.52	–	42.66	5.67
	Seq. ref.	**32.61**	**30.14**	**61.23**	**0.633**	55.21	**37.18**
4	OmniGen2	0.39	1.60	57.63	–	37.36	5.78
	UNO	19.60	17.35	**61.02**	–	56.01	14.65
	DreamO	0.19	0.75	57.39	–	42.34	4.63
	XVerse	2.67	2.63	57.71	–	36.54	3.17
	MOSAIC	2.46	1.99	57.34	–	37.73	2.61
	Seq. ref.	**20.14**	**19.48**	58.32	**0.611**	**56.12**	**35.52**
5	OmniGen2	0.08	0.43	57.38	–	36.85	2.49
	UNO	13.47	8.58	**59.26**	–	**56.40**	7.27
	DreamO	0.03	0.17	57.24	–	37.35	1.79
	XVerse	0.52	0.46	57.15	–	36.66	1.28
	MOSAIC	0.18	0.24	57.24	–	36.79	1.92
	Seq. ref.	**18.31**	**16.72**	57.98	**0.610**	55.21	**9.48**

Implementation Details. We construct the **CogCanvas** reference database from three disjoint subsets: Human (100 identities), Object (80 items), and Fashion (35 items). Each entity contains a name and reference images, each annotated with a unique description of its visual attributes.

All methods receive the same identity and object/fashion references; only our sequential pipeline has a dedicated background-image channel. For other methods, the background is specified in text, so BG-Sim is not reported and the background axis is not directly comparable. Face regions are parsed using facexlib BiSeNet to exclude neck and clothing from ID-Sim. DINOv2 features are extracted from cropped object/fashion regions. Gemma-4-e4b-it performs attribute extraction, prompt synthesis, three-query prompt verification, and Attr-VQA scoring (Appendices C.2 and C.5).

Computing Resources. All experiments use 4 NVIDIA A100 GPUs (40GB each) in parallel. Generating the full CogCanvas benchmark takes 10–12 hours per method, depending on inference steps and configuration. Computing the six metrics (ID-Sim, DINO-Sim, CLIP-T, BG-Sim, Aesth., and Attr-VQA) requires an additional $\sim$2 hours per method, with Gemma-4e4b-it-based Attr-VQA accounting for most of this time.

5.3 Multi-Subject Generation Results

Table 2 reports quantitative results grouped by subject count $N \in \{2, 3, 4, 5\}$. At $N = 2$, the sequential reference pipeline reaches ID-Sim 45.28 and Attr-VQA 53.41, compared with the strongest respective existing-method scores of 42.33 (XVerse) and 35.12 (UNO). These comparisons should be read alongside the different background-conditioning capabilities and runtime costs described above.

The main finding is a sharp decline in attribute fidelity as the number of subjects increases. At $N \geq 4$, four of five existing methods have DINO-Sim at most 2.63; UNO is the exception (17.35 at $N = 4$ and 8.58 at $N = 5$). The sequential pipeline retains DINO-Sim 19.48 and 16.72 at those group sizes. Attr-VQA also declines, from 35.52 at $N = 4$ to 9.48 at $N = 5$ for the pipeline, so it does not solve the full compositional task. Aesthetic scores alone conceal these compositional failures: UNO remains near 56 across group sizes even as its object/fashion score declines. The six axes should therefore be interpreted together. The underlying failure modes are dissected in Section 6.

6 DISCUSSION

Scaling Challenge. Table 2 reveals a consistent and severe degradation in performance as group size increases from $N = 2$ to $N = 5$ across all evaluated methods. For DINO-Sim, which measures object/fashion attribute preservation, four of five existing methods score at most 2.63 at $N \geq 4$. This indicates a severe loss of object/fashion fidelity in the evaluated methods beyond three subjects, with UNO as a partial exception. Attr-VQA scores are also low for most methods at $N = 5$. These patterns are consistent with cross-subject interference, although the present comparisons do not isolate a single causal mechanism.

Why Multiple Axes Matter. The table reports identity and text alignment separately because either score can look adequate while per-person binding fails; for example, XVerse reaches ID-Sim 42.33 at $N = 2$ but Attr-VQA 7.80. This divergence motivates reporting ID-Sim, DINO-Sim, and Attr-VQA together rather than using a single global score.

Human Validation of Attr-VQA. Seven annotators with backgrounds in computer vision and image generation independently answered Yes, No, or Unsure to questions about 46 generated images. The study covered 515 image–question pairs and yielded 2,869 valid votes after excluding Unsure responses. Human ground truth was the majority label for each question. At the image level, the proportion of Yes answers from Attr-VQA correlated with the human-majority proportion (Spearman $\rho = 0.645$, Pearson $r = 0.656$, Kendall $\tau_b = 0.506$). At the question level, the MLLM agreed with the human-majority label on 76.5% of 515 pairs (Cohen's $\kappa = 0.531$). These results support using Attr-VQA to compare broad trends, while its 57.64% mean Yes rate versus the human 51.00% shows that absolute scores can be optimistic. Human votes were evaluated against a majority label that includes each rater's own vote, so the human agreement figure is not a leave-one-rater-out estimate.

7 LIMITATIONS

CogCanvas has several limits. CelebA-HQ is an older celebrity source with restricted demographic breadth, and web-sourced objects or fashion may overlap generator training data. Thus absolute similarity scores may reflect memorization as well as reference grounding. The 29 backgrounds emphasize individually identifiable landmarks and natural scenes; generic interiors and physical or lighting coherence are outside the present test. The 196 five-person prompts support many relation-level checks, but do not by themselves establish statistical precision for every method comparison.

Gemma 4 is used in both prompt construction and Attr-VQA scoring, so shared blind spots remain possible. A human comparison on 46 images provides a limited validation, not a cross-family audit. Branded object and fashion imagery is represented through source URLs and metadata under a research-only release rather than redistributed as raw files; link rot may reduce long-term reproducibility. Sequential inpainting also increases inference cost with group size and does not eliminate interaction failures.

8 CONCLUSION

We introduced **CogCanvas**, a benchmark for evaluating multi-subject, reference-based image generation under realistic constraints. By combining a curated reference pool with structured interaction and position graph metadata, and a six-axis evaluation protocol, **CogCanvas** exposes systematic failure modes in current state-of-the-art methods that are invisible to existing single-subject or generic compositional benchmarks. Our results show that all current methods degrade substantially as group size increases from 2 to 5, with severe object/fashion fidelity loss for four of five existing methods at four or five subjects. We expect that **CogCanvas** will serve as a standardized tool for advancing research in multi-entity personalized image generation.

AI USE STATEMENT

Gemma 4 was used for reference attribute extraction, prompt synthesis and verification, and Attr-VQA evaluation as described in Sections 3.3 and 4. ChatGPT also assisted with language editing. The authors are responsible for the data, analyses, and text.

ETHICS STATEMENT

The benchmark includes public-person identities and links to third-party object and fashion images. We release source links and metadata rather than redistributing those images, and the research-only license prohibits impersonation, defamation, harassment, and non-consensual depiction. The identity pool has residual demographic bias; Appendix B discusses intended use and limits.

REPRODUCIBILITY STATEMENT

Sections 3.2–4 describe construction and scoring. Appendix C gives checkpoint settings, filtering thresholds, prompt verification, and evaluation details; Appendix D reports compute cost. The supplementary release provides prompts, graph annotations, manifests, and evaluation scripts.

REFERENCES

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

Black Forest Labs. FLUX.2 [klein] 9B. https://huggingface.co/black-forest-labs/FLUX.2-klein-9B, 2026. Hugging Face model repository, accessed 2026-05-07.

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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2506.20879.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

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. URL https://arxiv.org/abs/2506.21416. NeurIPS 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 (CVPR)*, pp. 4690–4699, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Yu Gao, Lixue Gong, Qiushan Guo, Xiaoxia Hou, Zhichao Lai, Fanshi Li, Liang Li, Xiaochen Lian, Chao Liao, Liyang Liu, et al. Seedream 3.0 technical report. *arXiv preprint arXiv:2504.11346*, 2025.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Shuhao Han, Jiacheng Chen, Siyuan Xu, Jian Wang, Liang Zhang, Wei-Wei Tu, Lian Liu, Yuhui Yuan, Xiaojie Jin, Jiashi Feng, and Xiaopeng Sun. EvalMuse-40K: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024. URL https://arxiv.org/abs/2412.18150.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 78723–78747, 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1710.10196.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Md Motaleb Hossen Manik and Ge Wang. Gemma 4, phi-4, and qwen3: Accuracy-efficiency tradeoffs in dense and moe reasoning language models. *arXiv preprint arXiv:2604.07035*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

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 (ICLR)*, 2025. URL https://arxiv.org/abs/2406.16855.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, 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 (CVPR)*, pp. 10684–10695, 2022.

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

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. LAION AI.

Dong She, Siming Fu, Mushui Liu, Qiaoqiao Jin, Hualiang Wang, Mu Liu, and Jidong Jiang. MOSAIC: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.01977.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

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*, 2025a.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. URL https://arxiv.org/abs/2504.02160.

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. URL https://arxiv.org/abs/2306.09341.

Yanze Wu, Yutong Feng, Difan Liu, Sabir Jariwalala, Zicheng Liu, Qiang Wen, Yuedong Yang, Ming-Hsuan Yang, and Chong Mou. DreamO: A unified framework for image customization, 2025c. URL https://arxiv.org/abs/2504.16915.

APPENDIX

A DATA AND CODE RELEASE

The anonymized dataset has been released at: https://huggingface.co/datasets/org-anonymous/CogCanvas.

The evaluation code has been released at: https://anonymous.4open.science/r/cogcanvas.

B SOCIETAL IMPACTS

Intended use. **CogCanvas** is designed as a research benchmark for evaluating multi-subject reference-based image generation. Its intended consumers are researchers and practitioners studying compositional generation, identity preservation, and controllable diffusion models. We release the benchmark to enable systematic measurement of the failure modes documented in Section 6, not to encourage deployment of identity-conditioned generation in production systems where consent and downstream use are not controlled.

Positive impacts. The benchmark exposes systematic weaknesses in current multi-subject generators that are invisible to single-subject benchmarks, providing a target for methodological progress. Including Vietnamese landmarks (Ha Long Bay, Dragon Bridge, Fansipan, Ba Na Hills, Hoi An Old Town) in the background pool broadens cultural representation in compositional generation benchmarks, which to date have been dominated by Western scenes. Because the benchmark is training-free with respect to the evaluated methods and ships full evaluation scripts, it lowers the barrier to entry for groups without access to large training compute.

Risks and mitigations. The most salient risk is that identity-conditioned generation can produce non-consensual likenesses or deepfakes. We mitigate this in three ways. First, all 100 identities are drawn from CelebA-HQ Karras et al. (2018), from which we select the identities; we add no new identities and do not redistribute the portrait images. Second, the dataset license explicitly prohibits use for impersonation, defamation, harassment, or any application that produces images depicting a real person without verifiable consent for that depiction. Third, we redistribute only the curated reference manifest (URLs and structured metadata) plus the prompt set; users must obtain CelebA-HQ separately under its existing license, which forecloses one-click bulk download of the identity pool.

Bias and representation. The identity pool is balanced along gender, age group, and ethnicity by a human-in-the-loop audit, but residual bias remains. CelebA-HQ over-represents European and East Asian celebrities relative to global demographics; our 100-identity selection is a best-effort rebalance within that source rather than a representative sample of any population. Object and fashion references are likewise biased toward items with strong web visibility, including a deliberate over-representation of culturally distinctive Vietnamese items (*e.g.*, Ao Dai variants) to support the geographic diversification of the background pool. Benchmark scores should therefore be read as comparative measurements across methods, not as evidence of any method's behavior on under-represented identities or items.

Evaluation reliance on a single MLLM family. Attr-VQA and prompt verification both rely on Gemma 4. This couples the benchmark to a single MLLM's blind spots: prompts the judge cannot evaluate reliably will receive systematically inflated or deflated scores. Cross-MLLM agreement audits are listed as an open item in Section 6 and we publish the full prompt and answer logs so that third parties can re-score the benchmark with alternative judges.

Memorization and disclosure. Because the identity pool overlaps with generator pretraining corpora, ID-Sim and DINO-Sim partly reflect memorization rather than reference grounding. We disclose this caveat in the main paper's Limitations section and re-emphasize it here: **CogCanvas** is a high-fidelity testbed for real-world compositional generation under naturally mixed-content inputs, not a probe of out-of-distribution generalization.

Release. We release the prompt set, structured interaction and position graphs, evaluation scripts, and a manifest pointing into CelebA-HQ and the original web sources for objects, fashion, and backgrounds. The license is research-only, with the explicit prohibitions described above. Any updated terms or known issues will be tracked at the project page.

C Implementation Details

This appendix expands the experimental setup in the main paper. We document (i) the curation pipeline, (ii) prompt synthesis, (iii) baselines and our training-free pipeline, and (iv) the evaluation pipeline, listing only the model checkpoints actually used to produce the numbers in Table 2.

C.1 Reference Curation

Person identities are drawn from CelebA-HQ Karras et al. (2018) and balanced along gender, age group, and ethnicity by a two-annotator audit. Object and fashion candidates are deduplicated with DINOv2 Oquab et al. (2024) on `[CLS]` features (cosine threshold 0.90), filtered by the LAION aesthetic predictor v2.5 Schuhmann (2022) at threshold 5.0, and screened by a CLIP ViT-L/14 Radford et al. (2021) text–image alignment check against the named query at threshold 0.25. A final manual review enforces the distinctiveness and image-quality criteria from Section 3.2. The 29 background scenes are individually reviewed for global recognizability, sufficient resolution (≥ 1024 px on the short side), and absence of incidental persons that would conflict with subject insertion.

C.2 Prompt Synthesis

Both attribute extraction and prompt synthesis use Gemma 4-e4b-it Manik & Wang (2026), decoded with `temperature`=0 under a JSON-schema-constrained generator. For every reference asset Gemma 4 emits structured fields—`age`, `gender`, `ethnicity` for identities; `size_category` and `afforded_interactions` for objects/fashion; `scene_type` and `landmark_name` for backgrounds—which condition the compositional sampler. Each candidate prompt is then passed through three Gemma 4 verification queries (interaction plausibility, layout feasibility, scene compatibility); the per-stage rejection rates are 9.4%, 6.1%, and 3.7%, and 1.2% of prompts require a human override. The final yield is 1,361 prompts.

C.3 Baselines

We benchmark five publicly released methods, including OmniGen2 Wu et al. (2025a), UNO Wu et al. (2025b), DreamO Wu et al. (2025c), XVerse Chen et al. (2025), and MOSAIC She et al. (2026), using their official Hugging Face checkpoints and the default sampling settings shipped with each release. Only the inputs (reference images, prompt, output resolution) are adapted to match **CogCanvas**; output resolution is 1024×1024 throughout and the seed is fixed for reproducibility. None of the five baselines exposes a dedicated channel for a background reference, so the background is provided as a textual mention in the prompt.

C.4 Training-Free Baseline

Our pipeline (Section 5.1) is implemented over FLUX 2 Klein Black Forest Labs (2026) as the diffusion backbone, with SAM Kirillov et al. (2023) for subject-region segmentation. *Stage 1* conditions FLUX 2 Klein on the background reference and on the compositional prompt with all identity names replaced by anonymous tokens, producing the base image at 1024×1024. *Stage 2* segments face and object regions for each subject slot. *Stage 3* inpaints subjects sequentially: for each $i = 1, \ldots, N$, FLUX 2 Klein runs masked diffusion conditioned on the subject's reference, then the canvas is re-segmented before the next subject. A single retry with seed $+1$ guards against detection drops; persistent failures contribute zeros to per-subject metrics.

C.5 Evaluation Pipeline

Identity (ID-Sim). Generated faces are embedded with ArcFace Deng et al. (2019) (IR-SE50 backbone, `model_ir_se50`) and scored by best-match cosine similarity against the prompt's ground-truth identities. Missing detections contribute 0.

Object/fashion similarity (DINO-Sim). Each object or fashion slot is localized in the generated image and embedded with DINOv2 Oquab et al. (2024); we report mean cosine similarity to the corresponding reference.

Background fidelity (BG-Sim). Both the generated image and the reference background are passed through SAM 3 Carion et al. (2025) with foreground subject classes suppressed; the remaining background regions are embedded with DINOv3 Siméoni et al. (2025) and aggregated by mean cosine similarity. We report BG-Sim only for the sequential reference pipeline because the five existing methods have no dedicated background-reference input.

Text alignment (CLIP-T) and aesthetics (Aesth.). CLIP-T uses CLIP ViT-L/14 Radford et al. (2021) on the prompt and full image; Aesth. uses the LAION aesthetic predictor v2.5 Schuhmann (2022).

Attribute VQA (Attr-VQA). The judge is Gemma 4-e4b-it Manik & Wang (2026), decoded with `temperature`=0 and `max_tokens`=2 against a binary Yes/No template. One query is issued per (subject, attribute) and per (subject-pair, interaction) entry of the structured graph; the final score is the mean accuracy across all questions, entities, and prompts.

D Computing Resources

Hardware. All experiments run on a single workstation with $4\times$ NVIDIA A100 (40 GB) GPUs, $2\times$ AMD EPYC 7763 CPUs (64 cores each), and 1 TB DDR4 RAM, connected via NVLink within node and PCIe Gen4 to disk.

Dataset curation cost. The curation pipeline (Section C.1–C.2) consumes approximately 38 GPU-hours: 12 h for DINOv2 deduplication and CLIP filtering of $\sim$15 k candidate object/fashion images, 18 h for the three-stage Gemma 4 prompt synthesis and verification of $\sim$4,000 candidate prompts (yielding 1,361 retained), and 8 h for human-in-the-loop demographic auditing of the identity pool. Human annotation contributed roughly 60 person-hours across two annotators.

Generation cost. Each baseline generates one image per prompt at 1024×1024 for all 1,361 prompts. End-to-end runtime ranges from 10–12 h per method on the 4-GPU node, depending on inference steps and conditioning overhead; our sequential-inpainting baseline costs 11.4 h because runtime grows linearly with N and dominates at $N=5$. The full benchmarking sweep across the six methods therefore consumes approximately 65 h of wall-clock time, or 260 GPU-hours.

Evaluation cost. Computing the six metrics for one method takes approximately 2 h; over six methods this totals 12 h (48 GPU-hours). Attr-VQA dominates this budget at roughly 78% of evaluation time because each prompt issues $O(N+\binom{N}{2})$ MLLM queries.

Reproducibility envelope. Because no method is trained or fine-tuned, reproducing every number in Table 2 requires only the released reference pool, the prompt set, the public baseline checkpoints, and the evaluation scripts; no GPU training time is needed. We will release inference and evaluation scripts that exercise the full pipeline on a single A100 in approximately one week of wall-clock time, or four days on a 4-GPU node.



Figure 5: **Qualitative comparison on multi-person samples (2, 3, and 4 persons) from CogCanvas.** The results illustrate the performance of our model compared to representative baselines in maintaining identity consistency and background alignment across increasing numbers of subjects.



Figure 6: **Qualitative comparison on five-identities samples from CogCanvas.** This figure demonstrates the model's capability in handling complex scenes with five distinct identities and specific background requirements.

E BENCHMARKING

E.1 QUALITATIVE ANALYSIS

Figure 5 traces the quantitative drops in Table 2 to a single root cause: OmniGen2, UNO, DreamO, XVerse, and MOSAIC are tuned for human-only or human–object personalization and expose no conditioning channel for the background reference that **CogCanvas** additionally requires. This failure mode is further exposed in the five-person samples at Figure 6: when multiple identity references must be jointly conditioned, OmniGen2, DreamO, XVerse, and MOSAIC all produce partial or misattributed outputs, whereas **CogCanvas** maintains per-subject fidelity throughout the scene.

Outputs therefore cluster into three recurring failure patterns: (i) *partial identity preservation*—faces match the target in coarse attributes (hairstyle, skin tone) but lose fine-grained features, consistent with low ID-Sim; (ii) *attribute leakage*—object/fashion items are dropped, swapped between subjects, or replaced by semantically related but visually distinct items (*e.g.*, a generic jacket in place of an "Ao Dai"), driving near-zero DINO-Sim at $N \geq 3$; and (iii) *background regression*—the specified scene is replaced by generic studio or outdoor backdrops regardless of whether the prompt requests Ha Long Bay, Dragon Bridge, or Fansipan. The sequential reference pipeline reduces some identity and binding failures by isolating subject conditioning, but its Attr-VQA score also declines at five subjects (Table 2). It therefore serves as a diagnostic control rather than a solution to all three failure modes.