MultiHuman-Testbench: Benchmarking Image Generation for Multiple Humans

Shubhankar Borse[§] **Seokeon Choi** **Sunghyun Park** **Jeongho Kim** **Shreya Kadambi**
Risheek Garrepalli **Sungrack Yun** **Munawar Hayat**[§] **Fatih Porikli**
Qualcomm AI Research*
[§]{sborse, mhayat}@qti.qualcomm.com

Abstract

Generation of images containing multiple humans, performing complex actions, while preserving their facial identities, is a significant challenge. A major factor contributing to this is the lack of a dedicated benchmark. To address this, we introduce MultiHuman-Testbench, a novel benchmark for rigorously evaluating generative models for multi-human generation. The benchmark comprises 1,800 samples, including carefully curated text prompts, describing a range of simple to complex human actions. These prompts are matched with a total of 5,550 unique human face images, sampled uniformly to ensure diversity across age, ethnic background, and gender. Alongside captions, we provide human-selected pose conditioning images which accurately match the prompt. We propose a multi-faceted evaluation suite employing four key metrics to quantify face count, ID similarity, prompt alignment, and action detection. We conduct a thorough evaluation of a diverse set of models, including zero-shot approaches and training-based methods, with and without regional priors. We also propose novel techniques to incorporate image and region isolation using human segmentation and Hungarian matching, significantly improving ID similarity. Our proposed benchmark and key findings provide valuable insights and a standardized tool for advancing research in multi-human image generation. The dataset and evaluation codes will be available at `https://github.com/Qualcomm-AI-research/MultiHuman-Testbench`.

1 Introduction

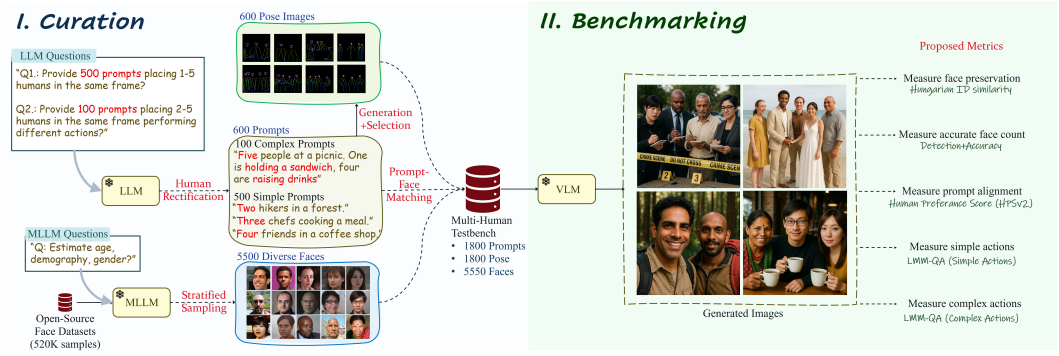


Figure 1: **MultiHuman Testbench**. Our MultiHuman Testbench consists of 5,550 IDs across 1,800 samples, including captions describing a scene with of 1-5 humans.

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

While current text-to-image diffusion models can generate high fidelity images, generating scenes featuring multiple humans (from provided reference images) performing text-described actions still remains a challenge. It requires simultaneously preserving visual characteristics of multiple subjects, accurately rendering their relative positions and interactions, and seamlessly integrating them into the synthesized background. However, current methods [44, 15, 43] frequently exhibit issues such as identity blending, generating the incorrect number of humans, or difficulties in composing the scene according to the text. To make the task easier, some works [21, 6] adopt regional priors as an input to the model, such as human poses, bounding boxes or segmentation masks. While this improves performance, it hinders usability as pose or mask information might not be readily available.

A major challenge in multi-human generation is the lack of a comprehensive and standardized benchmark, along with well defined metrics. Existing benchmarks typically focus on single-subject fidelity [12, 4], general text-to-image quality [11], or multi-object compositional tasks [24, 14]. However, none of the currently available open benchmarks address the added complexity of generating multiple distinct humans. To address this issue, we introduce MultiHuman-Testbench, a novel and challenging benchmark. It is built upon a dataset of 1,800 samples, which include carefully crafted text prompts describing scenes with 1 to 5 humans, paired with 5,550 reference human faces, sampled from open-source datasets. We ensure diversity in age, ethnicity, and gender. As many current works rely on regional priors in multi-human scenes [21, 6], we provide pose conditioning images. Additionally, we propose a multi-faceted evaluation framework designed to capture the nuances of multi-human generation. We propose four complementary metrics: Count Accuracy, Hungarian ID similarity, Human Preference Score, Multimodal LLM (MLLM) question-answering to probe the correctness of simple and complex actions. The proposed testbench has four different tasks: 1) Reference-based Multi-Human Generation in the wild. 2) Reference-based Multi-Human Generation with Regional Priors. 3) ID-Consistent Multi-Human Generation without Reference Images. 4) Text-to-Image Multi-Human Generation. We benchmark current models and identify key areas for improvement. Overall, most methods without regional priors struggle in generating the correct number of people. While proprietery models such as GPT-Image-1 generates plausible images, it lacks preserving facial features and has poor ID retention. We also study biases in current models, in terms of gender, age, status, and ethnicity.

Reference-based Multi-Human Generation in the wild, is the most challenging and least restrictive task in our testbench. We propose new techniques (Sec. 3) to adapt current methods for improving their performance this task. Specifically, for unified multi-modal architectures [44, 30], we propose a method to isolate the reference images to impact only a specific region within the latent space. To match each reference image to regions, we propose an implicit Hungarian matching guided by human segmentation. Our method enhances the ability to maintain individual identities, reducing subject leakage and improving ID similarity. We extend our proposed techniques to two models, OmniGen [44] and IR-Diffusion [13], resulting in our proposed MH-OmniGen and MH-IR-Diffusion.

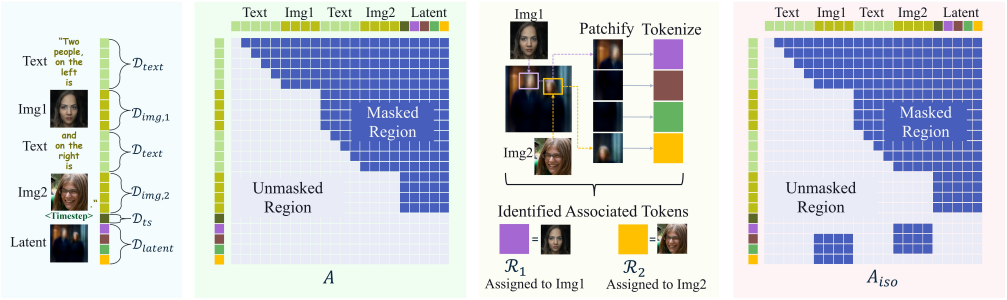
In summary, our contributions are:

- Introduction of a novel benchmark for multi-human ID image generation, featuring diverse subjects, text, and pose conditioning.
- A comprehensive evaluation suite designed to assess multi-human generation fidelity, including people count accuracy, ID similarity, text-alignment, and MLLM-based assessment.
- An extensive empirical evaluation and thorough analysis of 30 state-of-the-art zero-shot and training-based generative methods on four different tasks.
- A novel training-free enhancement for existing multi-human generation methods, utilizing regional isolation and matching for improved identity and compositional control.

2 Related Work

Native Text-to-Image models: Multiple diffusion based models have been proposed recently [32, 36, 37, 10, 5]. These models exhibit excellent text-to-image generation ability and can be used as base models for generating multiple humans.

Multi-human Generation with native Text-to-Image models: To generate ID-consistent or subject driven images with text-to-image models, recent works employ auxiliary models such as IP-Adapter [46] or ControlNet [48]. There are also tuning-based approaches which exist such as LoRA [16] or



a. Token space b. Attention Mask (Unified) c. Regional Association d. Unified Regional Isolation

Figure 2: **Regional Isolation for Unified Architectures**. The updates to the attention mask for regional isolation are illustrated in the differences between Fig.b and Fig.d.

MudI [39] for this purpose. These methods typically fall short on multi-human generation in the wild, without any regional priors.

ID-Consistent Multi-Human Generation without reference images: Methods such as Consistory [39], DreamStory [14], IR-Diffusion [13] and StoryDiffusion [49] have recently gained popularity in ID-consistent multi-subject generation without reference images. These methods generate human faces and use these faces to generate multiple images for tasks like storytelling.

Multi-Object Generation: Recent approaches e.g., MS-Diffusion [41], MIP-Adapter [17], Lambda-Eclipse [31], have shown significant performance gain for incorporating multiple objects in the same scene. These can include daily items and in some cases, pets such as dogs and cats. However, they struggle to adapt to multi-human generation as zero-shot ID-preservation is a highly challenging task.

Multi-human Generation with native Multi-Modal models: Unified multimodal models, such as OmniGen [44], Show-O [45], OneDiffusion [26] ACE++ [29], GPT-Image-1 [30] and JanusFlow [28] process the text and vision via same transformer backbone and have shown promises for subject-driven generation. These methods input both the reference images and text prompt in a unified token space, removing the need for additional auxiliary task-specific networks such as IP-Adapter/ControlNet. Omnigen [44] was further tuned for ID-preservation. Our evaluations show that among all open-source models, Omnigen produces best results.

Regional Isolation: For networks generating images using simply text inputs, recent works such as IR-Diffusion [13] and InstantFamily [21] have proposed methods such as image isolation and repositional attention, which successfully localize multiple humans in the scene by isolating them from each other and mapping them to separate regions in the image latent. These methods have shown great promise in reducing leakage between multiple human identities.

3 Proposed Approach: Enhancing Existing Methods

Reference-based Multi-human generation in the wild (Task 1 in 5.1) is a highly challenging problem. It requires to preserve input identity while rendering the complete scene with the correct number of humans performing a described action. Using insights from benchmarking current approaches in Section 5, we observe several limitations, including identity blending or missing identities. To tackle these issues, we propose two techniques: Unified Regional Isolation and Implicit Regional Assignment, that can be flexibly incorporated with existing methods to enhance their quality.

Unified Regional Isolation: Motivated by [21, 13] for T2I architectures, we develop a regional isolation masking strategy to tackle the limitations relating to identity blending and missing, tailored for unifed models such as OmniGen [44].

Consider the token space for a unified multimodal model represented in Figure 2 (a). Let L be the total sequence length, and let $i, j \in \{1, \ldots, L\}$ be the indices for query and key/value tokens. We define disjoint sets of indices for each token type: $\mathcal{D}_{\text{text}}$, $\mathcal{D}_{\text{img}}$, $\mathcal{D}_{\text{ts}}$(for timestep), and $\mathcal{D}_{\text{latent}}$. Using the setup from OmniGen [44] and Show-o [45], the self-attention mask $\mathbf{A}$ ($L \times L$) is constructed based on the type of the query token i: causal attention for text queries, and bidirectional attention for non-text queries (image, timestep, latent). This is represented as:

$$A_{ij} = \begin{cases} 1 & \text{if } i \in \mathcal{D}_{\text{text}} \text{ and } j \leq i \quad \text{(text query: causal)} \\ 1 & \text{if } i \notin \mathcal{D}_{\text{text}} \quad \text{(non-text query: bidirectional)} \\ 0 & \text{otherwise} \end{cases}$$

Consider the tokens in $\mathcal{D}_{\text{img}}$ are derived from N distinct original input images, $\{I_1, \ldots, I_N\}$. For each image I_k ($k = 1, \ldots, N$), let $\mathcal{D}_{\text{img},k} \subseteq \mathcal{D}_{\text{img}}$ be the set of sequence indices corresponding to its derived tokens. These sets partition $\mathcal{D}_{\text{img}}$. This is represented in Figure 2 for ($k = 1, 2$). For each reference image I_k, consider that we find a region of interest (ROI) set $\mathcal{R}_k \subseteq \mathcal{D}_{\text{latent}}$. Now, we construct a new attention mask $\mathbf{A_{iso}}$ ($L \times L$) such that it isolates the images I_k to only the specific region $\mathcal{R}_k$ within the latent. Hence, our proposed attention mask is computed as:

$$A_{iso,ij} = \begin{cases} 1 & \text{if } i \in \mathcal{D}_{\text{text}} \text{ and } j \leq i \quad \text{(text query: causal)} \\ 1 & \text{if } i \in \mathcal{D}_{\text{img}} \text{ and } (j \notin \mathcal{D}_{\text{latent}} \text{ or } j \in \mathcal{R}_k \text{ where } i \in \mathcal{D}_{\text{img},k}) \quad \text{(image query: ROI attention)} \\ 1 & \text{if } i \in \mathcal{D}_{\text{ts}} \cup \mathcal{D}_{\text{latent}} \quad \text{(timestep/latent query: bidirectional)} \\ 0 & \text{otherwise.} \end{cases}$$

Implicit Region Assignment: To construct the attention mask $\mathbf{A_{iso}}$, we need region of interest for every image $\mathcal{R}_k$. This can be done explicitly as in recent methods InstantFamily [21] and Regional Prompting [6], or using a regional prior (pose conditioning or bounding boxes). However, this severely hinders usability, as the users might not want to seek for a multi-human pose image resembling the one which they wish to generate. Hence, to facilitate the generation of multi-human images in the wild, we propose an implicit region assignment strategy that utilize intermediate attention scores and Hungarian matching to assign each reference image to a selected region-of-interest.

Below, we discuss adaptation of our proposed techniques for different models including Omingen [44] and IR-Diffusion [13]. See Appendix C for further details and algorithms.

MH-Omnigen: To find optimal regions for architectures such as Omnigen [44] which have a unified token space for text and reference images, we probe the backbone transformer model at an intermediate timestep. The self-attention maps in the backbone transformer model provide information for the regional overlap for reference images, and the segmentation masks of the intermediate latents provide regional information for each generated person. We perform hungarian matching to find reference inputs with the maximum self-attention region, to eventually find $\mathcal{R}_k$.

MH-IR-Diffusion: In the case for IR-Diffusion [13], the region-of-interest is defined by the models initially generated images. Similar to the original work, we use a segmentation model, SAM2 [34] to generate the region proposals for generated faces. Next, we compute Arcface similarity between generated faces and reference faces to find the best match, and utilize hungarian matching to assign segments of the matched faces as regions $\mathcal{R}_k$.

Our experiments in Sec. 5 show that Regional Isolation and Implicit Assignment are training-free plug-and-play methods which can effectively improve different baselines. Due to the implicit matching of identities and localization, we get improved ID similarity with reduced subject blending artifacts.

4 MultiHuman Testbench

Below we elaborate the process of curation of our proposed testbench, and discuss different metrics.

4.1 Image Selection

We curate images using three existing large-scale image datasets, FFHQ [20], SFHQ [1] and CelebaHQ [19], which initially contained approximately 520k samples. These datasets underwent a multi-stage filtering process, where

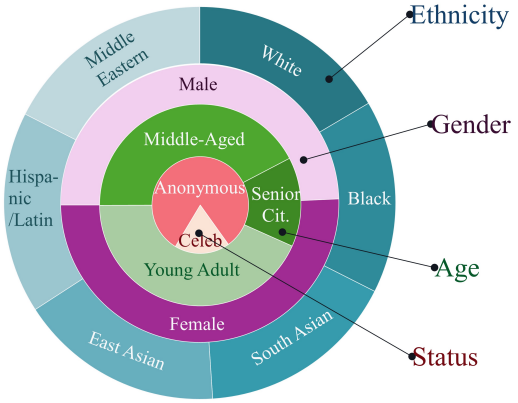


Figure 3: **Data distribution** among four major attributes: Ethnicity, Age, and Gender, Status. See Appendix B for details.

Figure 4: **Wordcloud**. The graphic shows words from our caption space.

initially we removed human IDs deemed non-
"identifiable", using MLLM [27] VQA. We prompt the MLLM with the question "Is the person's face identifiable AND unobstructed?" The images with negative response are filtered out. Subsequently, we identify multi-face images using face detection and eliminate them. These steps reduced the dataset size to 94k distinct human face images. For annotation, we employed Gemini Flash 2.0 [7] to classify each of these images based on three predefined attributes: estimated age-bracket (one of 16-35, 35-60, 60+), estimated demographic (sorted into 6 categories: Caucasian/White, Black/Native African, South-Asian, East-Asian, Hispanic/Latin, Middle Eastern/North African), and Estimated Gender assigned at birth (divided into 2 categories: Male, Female). Our dataset does not contain any images of minor subjects. Following this annotation, a non-biased test set consisting of 5,550 images was generated from the curated and labeled dataset using a stratified sampling approach to ensure representative distribution across the various age, ethnicity, and gender buckets. See Figure 3 for the final distribution. For ethnicity and gender, we provided a target uniform distribution. For age-bracket, we set a target percentage of 15%, using insights from PopulationPyramid [33].

4.2 Prompt Curation and Assignment

We aim to create a diverse set of prompts and begin by using Gemini Flash 2.0 [7] to generate an initial set of 100 prompts, each describing five people performing the same action. This set was then recontextualized to create variations for scenarios involving one to four people, forming the core of our "simple prompt" set. In parallel, we curated a separate collection of 25 prompts specifically designed to depict multiple people performing different actions within the same image, which were also recontextualized for scenarios involving two to four people, resulting in our "complex prompt" set. Our prompt set comprises the combination of these two collections, totaling 125 distinct prompts. The simple prompts cover scenarios with one to five people, while the complex prompts are intended for two to five people. To build MultiHuman Testbench, we generated testing samples by sampling three random iterations of human IDs for each prompt, resulting in a total of 1,800 unique testing samples. Each prompt underwent careful human revision after multiple iterations of generation with various subject-driven generation models to ensure the final set's quality and suitability. The wordcloud for our final prompt action space is visible in Figure 4. To populate the 1,800 testing samples, each requiring a specific human ID from the 5,550-image test set, a distributed sampling approach was employed to assign the 5,550 human IDs to the prompt iterations.

4.3 Pose Image Estimation

Pose images act as Regional Priors for benchmarking in Task3 (see Sec. 5.1). Hence, to obtain suitable pose information for the MultiHuman Testbench, we generate pose from two sources: a) the best results for each prompt using our outputs from Tasks 1 and 4, and b) a Text-to-Pose generation [3] model. Using resulting Images from Tasks 1 and 4, we find all images with the best Count and Action metrics for each prompt, and filter the ones which fall below a specific threshold (i.e. below 0.97 for action similarity, and below 1 for Count Accuracy). For prompts which ar filtered out due to poor overall scores, we obtain pose information using Text-to-pose [3]. We generate 20 distinct pose samples for each prompt. A crucial human selection step was then performed, where reviewers carefully evaluated each generated option to identify the single best pose for every prompt in our dataset. This rigorous manual filtering was necessitated by the inherent limitations of the text-to-pose generation model (along with limitations of Task 1 and Task 4), which occasionally demonstrate susceptibility to generating erroneous poses. The results presented in Appendix E provide strong evidence for the effectiveness of our provided pose pairs, yielding significantly improved results across most metrics.

4.4 Multi-View Image Generation

For each human ID in the curated dataset, we utilized the PuLID-Flux [12] to generate a collection of five distinct images. We prompt the model to render the person's identity in diverse contexts, capturing various perspectives such as full-portrait and side views, and placing them in differing environments. The intent behind these multiple generated images per ID is to gather comprehensive training data to improve performance on tuning-based multi-human generation models.

4.5 Metrics

To evaluate multi-human image generation, our benchmark proposes a suite of metrics specifically designed to capture various critical aspects of the generated output.

Hungarian ID Similarity. We propose an ID similarity metric using ArcFace embeddings [9]. To match input and generated IDs in the multi-human setting, we use cosine similarity of Arcface embeddings, and use the hungarian algorithm [23] to match each face while maximizing cost. The Hungarian ID similarity for a given image is thus the average matched ID similarity.

Consider a set of N input face images, indexed by $i = 1, \ldots, N$, and a set of M output face detections in the generated image, indexed by $j = 1, \ldots, M$. Consider Arcface embeddings for input images $\mathbf{F}^{\text{ref}} = \{\mathbf{f}_i^{\text{ref}} \mid i = 1, \ldots, N\}$, where $\mathbf{f}_i^{\text{ref}} \in \mathbb{R}^d$ and for generated faces $\mathbf{F}^{\text{gen}} = \{\mathbf{f}_j^{\text{gen}} \mid j = 1, \ldots, M\}$, where $\mathbf{f}_j^{\text{gen}} \in \mathbb{R}^d$. Here, d is the dimensionality of the feature space. Next, we define the similarity s_{ij} between reference face i and generated face j using cosine similarity:

$$s_{ij} = cosSim(\mathbf{f}_i^{\text{ref}}, \mathbf{f}_j^{\text{gen}}) = \frac{(\mathbf{f}_i^{\text{ref}})^\top \mathbf{f}_j^{\text{gen}}}{\|\mathbf{f}_i^{\text{ref}}\|_2 \|\mathbf{f}_j^{\text{gen}}\|_2}$$

We form an $N \times M$ similarity matrix $\mathbf{S}$, where $\mathbf{S}_{ij} = s_{ij}$. Since the Hungarian algorithm finds a minimum cost assignment, we define the cost c_{ij} as the negative similarity, $c_{ij} = -s_{ij}$. Using the Hungarian algorithm, we find a binary assignment matrix $\mathbf{X}$ ($X_{ij} = 1$ if matched, 0 otherwise). For each reference input i, if reference i is matched to a generated face j (i.e., $\sum_{k=1}^{M} X_{ik} = 1$), its contribution to the ID metric is the similarity s_{ij}/N for the matched j. If reference i is not matched to any generated face (i.e., $\sum_{k=1}^{M} X_{ik} = 0$), its contribution to the ID metric is 0. Hence, the average similarity over all N reference inputs, denoted S_{id}, is denoted as: $S_{id} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} X_{ij} s_{ij}$. Our proposed Hungarian ID similarity metric objectively evaluates the model's effectiveness in maintaining consistent and uniquely recognizable identities across different generations. Further, the proposed metric penalizes for subject/ID mixing.

Count Accuracy. Next, we assess the accuracy of the generated people count. This verifies the model's ability to precisely adhere to the numerical specification in the prompt. We use a face detection model [8] to count detected human faces in the generated image. Hence, as N is the number of reference images and M is the number of generated faces, the count accuracy is $S_{\text{count}} = \delta_{MN}$, where δ is the kronecker delta [2] function.

Quality/Prompt Alignment. Third, text alignment for overall scene consistency is evaluated using the HPSv2 score [42], S_{hps}. This metric goes beyond individual elements to measure how well the entire generated image corresponds to the textual description of the scene, ensuring that contexts, environments, and overall narrative specified in the prompt are accurately reflected.

MLLM Action QA. Fourth, to probe the correctness of simple and complex actions and interactions among multiple individuals, we utilize Multimodal Large Language Model (MLLM) question-answering. This approach allows for a deeper semantic evaluation by querying the MLLM about specific details, activities, and relationships depicted in the generated image, thereby assessing challenging compositional aspects. We propose to report the average separately for simple actions (Action-S) and complex actions (Action-C), as they provide deeper meaning. To generate the questions, we probe Gemini-Flash [7] to extract actions from each text prompt, and re-contextualize these into questions. For instance, assume the prompt is "Five people caroling during winter: among them, two people are holding song books, and three people are singing". For this prompt, the questions which are generated to rank complex actions are as follows: "Q1: Are two people in this image holding song books? Choices: 1(No), 10(Yes), 5(Partially)? Q2: Are the people in this image caroling? Choices: 1(No), 10(Yes), 5(Partially)?". Hence, the final (Action-C) score is the average score.

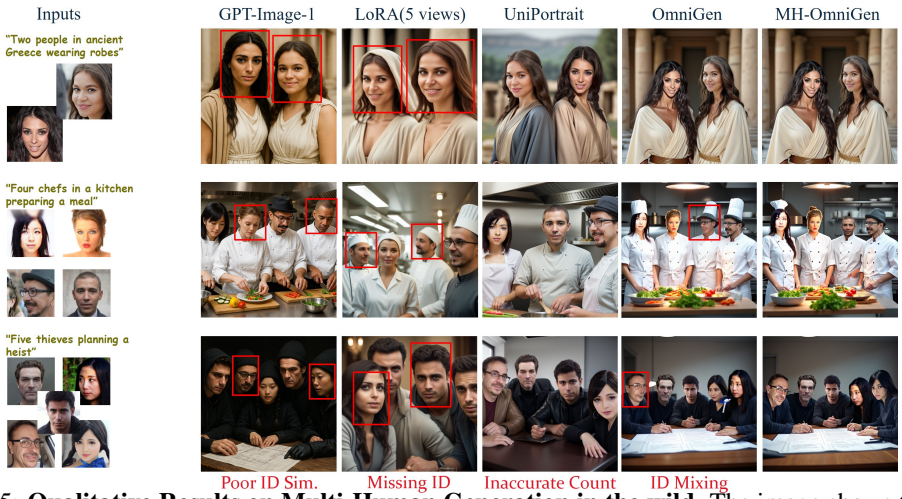


Figure 5: **Qualitative Results on Multi-Human Generation in the wild**. The image shows the best performing methods: UniPortrait, LoRA, GPT-Image-1, OmniGen and MH-OmniGen.

Unified Evaluation Metric. While individual metrics provide granular insights, practical model comparison benefits from a unified score. We propose a composite metric that integrates identity fidelity and prompt-image alignment. Specifically, we compute the geometric mean of (i) Hungarian ID similarity S_{id} and (ii) a weighted aggregate of alignment metrics (HPS, Action-S, Action-C, and Count Accuracy). The geometric mean ensures that poor performance in either dimension significantly reduces the overall score, reflecting the interdependence of identity preservation and semantic correctness.

Let:

$$S_{\text{align}} = \frac{S_{\text{hps}} + S_{\text{act}} + S_{\text{count}}}{3}.$$

To emphasize alignment, we apply a quadratic weighting to S_{align} within the geometric mean. The unified metric S_{U} is:

$$S_{\text{U}} = \left(S_{\text{id}} \times (S_{\text{align}})^2\right)^{\frac{1}{3}}.$$

This formulation ensures that if either identity preservation or alignment is weak, the overall score remains low, promoting balanced optimization across all aspects of multi-human image generation.

5 Benchmarking

We benchmark models under four settings: 1). Reference-based Multi-Human Generation in the wild, 2). Reference-based Multi-Human Generation with Regional Priors, 3). Story Consistent Multi-Human generation without reference images and 4). Text-to-Image multi-human generation without reference images. We use tuning-based methods such as LoRA [16] and MuDI [18] on SDXL [32]. For each prompt, we tune a single LoRA for all concepts, using the method in [18]. Next, we benchmark methods trained for general multi-subject generation, MS-Diffusion [41], MIP-Adapter [17] and Lambda-Eclipse [31]. For multi-human generation, we evaluate UniPortrait [15], RectifID [38], Fastcomposer [43], OmniGen [44], Flux-Kontext [25] and GPT-Image-1 [30]. We also benchmark methods which require explicit regional priors such as Regional-Prompting [6] with PuLID [12], and OMG [22] with InstantID [40]. Furthermore, we evaluate story-based(reference-free) diffusion models, Consistory [39], DreamStory [14] and IR-Diffusion [47] on our proposed benchmark. Finally, we evaluate native text-to-image models, SD-1.5 [36], RV-1.5 [37], SDXL [32], SD3.5 [10], Flux [25] and OmniGen [44], for generating accurate number of humans. All implementation details and hyperparameters are provided in Appendix D.

5.1 Results

Task 1. Reference-based Multi-Human Generation in the Wild:

Our results for task 1 are summarized in Table 1. We evaluate the performance of four different types of models: Proprietary, Tuning-based, Multi-Object Tuning-Free and Multi-Human Tuning-

Free methods. From the scores, we find that the performance of each method is significantly influenced by the backbone model it builds upon. On average, **Multi-object Tuning-Free methods** perform worse compared to other approaches. This is because generating humans and keeping their likeness intact is a significantly challenging problem, compared to the objects the methods have been trained on. Next, **Tuning-based methods** MuDI [18] and LoRA [16] perform slightly better, but are significantly bounded by the base architecture SDXL. As observed, training with 5 views generated from PuLID [12] performs better than a single view. Moving to **Multi-Human Tuning-free** approaches, we can observe that UniPortrait [15], built on RV1.5, significantly outperforms other SD1.5-based approaches such as RectifID [38] and FastComposer [43] across all metrics. However, unified multi-modal models, OmniGen [44] and GPT-Image-1 [30], perform significantly better than the rest. Notably, our proposed MH-OmniGen consistently outperforms its predecessor across four of the five metrics. We observe a **5.1** point difference in Multi-ID and **4.1** point difference in action similarity. This validates the effectiveness of our Unified Regional Isolation and Implicit Assignment method. Finally, among all evaluated methods, GPT-4o (via GPT-Image-1) achieves the highest overall performance in count accuracy, HPS, and action-based metrics. However, its performance in ID similarity is notably weaker (**25.7** points) than MH-OmniGen. This is due to the fact that GPT hallucinates features on humans, and in many cases isn't able to effectively maintain the identity of the person. When considering the Unified metric, which balances all aspects of performance, MH-OmniGen achieves the best score (**61.6**), followed by OmniGen (**59.2**) and GPT-Image-1 (54.3), demonstrating that our proposed method achieves the strongest overall balanced performance across all dimensions. **Overall**, we want to stress that **None** of the methods perform consistently well at a high standard for this task in terms of visual quality. Within the open-source methods, **None** of the models can consistently generate images with a high Action-C or Count score. There is significant scope for improvement in this setting.

Qualitative Results: In Figure 5, we show visual Results for the best models performing Task 1. As observed in this image, GPT-1 isn't able to effectively maintain human ID, owing to poor scores. On the other hand, Uniportrait generates good results but often with the inaccurate number of humans. OmniGen results have artifacts related to ID mixing, which are considerably repaired in MH-OmniGen. However, OmniGen-based methods tend to "copy" human faces. It is important to note that these are some of the better looking images for each method. We share more visual results in the Appendix, highlighting a heavy scope for improvement.

	Backbone	Model	Metrics					
			Count	Multi-ID	HPS	Action-S	Action-C	Unified
Task 1: Reference-based Multi-Human Generation in the Wild								
Proprietary	GPT-4o	GPT-Image-1	87.9	28.8	30.3	97.0	91.1	54.3
Tuning-Based	SDXL	LoRA(1 view)	47.3	20.2	25.3	61.0	55.4	36.2
		LoRA(5 views)	52.6	22.0	25.9	73.0	72.9	41.0
		MuDI(1 view)	48.1	23.6	24.8	64.0	51.5	37.7
		MuDI(5 views)	53.9	24.6	25.6	67.3	71.5	42.3
Multi-Object Tuning-Free	SDXL	IP-Adapter	34.3	9.3	23.2	49.6	46.9	16.3
		MIP-Adapter	39.2	11.9	24.0	57.6	53.7	19.2
	Kand2.2	Lamda-Eclipse	53.3	12.5	23.4	56.1	50.8	23.1
Multi-Human Tuning-Free	RV1.5	UniPortrait	58.5	44.2	25.9	76.2	67.2	51.7
	SD1.5	RectifID	37.8	18.6	24.8	67.3	68.2	33.8
		Fastcomposer	31.2	12.2	21.7	48.9	41.2	20.2
	Phi-3	OmniGen	**60.5**	49.4	26.2	87.5	71.3	59.2
		MH-Omnigen	60.3	**54.5**	**26.3**	**91.6**	**72.9**	**61.6**

Table 1: Multi-Human Generation with Reference Images in the wild.

Task 2. Reference-based Multi-Human Generation with Regional Priors:

Next, we evaluate methods for Task 2, which focuses on reference-based multi-human generation with regional priors. Table 2 shows results for Tuning-based, Multi-Object Tuning-Free, and Multi-Human Tuning-Free methods, all leveraging pose or box priors. As observed, the introduction of our provided human-rectified pose priors significantly improves quantitative metrics, particularly Count accuracy, compared to Task 1. We also observe a general increase in Action scores. For instance, MIP-Adapter [17] shows significantly higher Count accuracy with pose priors. Within the Multi-Human Tuning-Free group, we observe varied performance across metrics depending on the backbone and specific prior usage. RectifID [38] achieves the highest Count accuracy (**90.1**), while

OMG-InstantID [22] (SDXL) excels in HPS (**27.2**) and Action scores (**90.4**, **78.9**), and Regional-PuLID [6] (Flux) shows the strongest Multi-ID retention (**50.7**). Flux-Kontext, also based on the Flux backbone, demonstrates strong performance in Action scores (80.9, **79.8**), achieving the highest Action-C score among all methods, while maintaining competitive Count accuracy (76.8) and HPS (26.9). OmniGen demonstrates competitive performance in Task 2, maintaining strong HPS (**27.4**) and Action scores (**86.2**) and decent Multi-ID (**48.2**) when incorporating pose priors. Overall, Task 2 results highlight the significant benefit of regional guidance for key metrics like count, while demonstrating that achieving high performance across all aspects (like ID fidelity, action, and overall quality) remains a challenge. This is due to the fact that different methods are strong in different areas. When considering the Unified metric, which balances all aspects of performance, UniPortrait and OmniGen achieve the highest scores (**62.5**), followed by Regional-PuLID (56.4), Flux-Kontext (55.7), and RectifID (55.4), demonstrating that multi-human tuning-free methods generally outperform tuning-based and multi-object approaches in overall balanced performance.

	Backbone	Model	Regional Conditioning	Metrics					
				Count	Multi-ID	HPS	Action-S	Action-C	Unified
Task 2: Reference-based Multi-Human Generation with Regional Priors									
Tuning-based	SDXL	LoRA(1 view)	Pose	85.3	17.3	26.1	73.6	78.0	40.3
		LoRA(5 views)		89.6	21.4	26.0	77.7	73.6	44.1
Multi-Object Tuning-Free	SDXL	MIP-Adapter	Pose	81.5	14.1	25.0	69.2	67.2	36.9
Multi-Human Tuning-Free	Flux	Regional-PuLID	Boxes	67.4	**50.7**	26.1	74.1	68.0	56.4
	Flux	Flux-Kontext	Boxes	76.8	39.2	26.9	80.9	**79.8**	55.7
	RV1.5	UniPortrait	Pose	78.3	49.2	26.3	88.2	78.1	**62.5**
	SD1.5	RectifID	Pose	**90.1**	26.4	25.7	78.7	73.5	55.4
	SDXL	OMG-InstantID	Pose	71.2	32.6	27.2	**90.4**	78.9	54.6
	Phi-3	OmniGen	Pose	77.2	48.2	**27.4**	86.2	75.3	**62.5**

Table 2: Multi-Human Generation with Reference Images, with regional priors

Task 3. ID-Consistent Multi Human Generation without reference images:

Table 3 displays results on Task 3. We benchmarked four approaches. The performance varies across metrics, with ConsiStory and DreamStory showing lower accuracy in Count and ID-Similarity compared to IR-Diffusion and MH-IR-Diffusion. Notably, MH-IR-Diffusion achieved the highest scores in both Count (**62.6**) and Multi-ID (**33.3**). Due to the process of masking and hungarian assignment, the model is successfully able to preserve ID information while keeping the remaining metrics stable. While IR-Diffusion led slightly in Action-S (**86.3**), and all models performed similarly on HPS. We observe for the Complex Action metrics, that the performance reduces slightly as ID similarity improves. This is due to the fact that the results are closer to original model generation, as lesser ID's have been matched. Overall, we see a significant scope of improvement for every method in this list, due to poor performance on all metrics.

Backbone	Model	Resolution	Metrics				
			Count	Multi-ID	HPS	Action-S	Action-C
Task 3: ID-Consistent Multi-Human Generation without Reference Images							
Playground-v2.5	ConsiStory	768×1280	44.6	16.2	28.0	84.1	**71.9**
	DreamStory		45.0	19.7	28.2	84.8	71.8
	IR-Diffusion		62.4	27.6	29.4	**86.3**	71.8
	MH-IR-Diffusion		**62.6**	**33.3**	**29.2**	85.9	71.3

Table 3: Multi-Human Generation without Reference Images

Task 4. Text-to-image Multi-Human Generation:

For Task 4, we evaluate the overall effectiveness of text-to-image methods on generating humans with accurate count performing text-described simple and complex action. This is mainly to motivate the selection of a suitable base architecture for follow-up methods to build their solutions on. Our results are summarized in Table 4. Across all methods we evaluated, Flux, SD3.5 and OmniGen perform best, given the fact that they are larger models and have been trained on richer data. Notably, Flux produces images with consistent Count accuracy over 3, 4, 5 humans, compared to the other methods which show a steeper drop in performance after increasing the number of humans. SD3.5 is highly competitive in generating the correct actions (simple and complex), and OmniGen produces

the best HPS. Overall, however, there is significant scope for improvement in terms of human count, as the best score (**63.9**) is still quite low.

Model	Person Count				Prompt Alignment		
	3-Person	4-Person	5-Person	Avg(1-5)	HPS	Action-S	Action-C
Task 4: Text-to-Image Multi-Human Generation							
SD-1.5	25.0	12.8	7.4	26.6	24.8	84.5	73.3
RV-1.5	52.5	20.3	11.5	43.5	26.9	88.3	76.0
SDXL	44.0	30.7	23.5	43.3	26.9	87.9	79.1
SD3.5	61.0	<u>45.6</u>	28.8	<u>56.1</u>	27.8	**95.7**	**85.0**
OmniGen	<u>64.0</u>	29.0	<u>33.2</u>	53.8	**28.7**	<u>93.2</u>	82.5
Flux-Dev	**66.9**	**57.0**	**46.4**	**63.9**	<u>28.2</u>	92.6	<u>83.0</u>

Table 4: Benchmaking Foundational Text-to-Image models on generating multiple people.

5.2 Scope for Improvement

From the analysis in this Section and in Appendix E,F, we uncover several limitations of current approaches performing Multi-Human Generation. **First**, we notice that without regional priors, the open-source models performing Tasks 1, 2, 4 are lacking in terms of the person count and complex actions. Essentially, this means that multi-human generation is significantly hindered because the base model itself (from task 4) isn't able to generate an accurate number of human faces in a scene within the text described action. **Second**, we notice that none of the methods (with OR without regional priors) consistently pass the eye test in generating images of the correct number of people while maintaining a high ID similarity, and sufficient action scores. This balance is essential for our proposed task, and remains an open challenge. **Third**, in Appendix E, we observe that several methods contain implicit biases over age, racial profile and gender. After uncovering these biases, we hope that the community strives to reduce them using insights from our benchmark.

6 Conclusion

This paper introduces MultiHuman-Testbench, the first comprehensive benchmark for subject-driven multi-human image generation. We contribute a carefully curated dataset of 1,800 testing samples with balanced demographic representation, a multi-faceted suite of metrics capturing count accuracy, identity preservation, visual quality, and action consistency. We also propose training-free enhancements to unified human generation models through Unified Regional Isolation and Implicit Assignment. Through extensive evaluation of approximately 30 models across four distinct tasks, we reveal that current state-of-the-art methods exhibit significant limitations. Even the best-performing models struggle with accurate human counts and preserving individual identities without subject blending artifacts. Our analysis highlights substantial opportunities for future research in achieving robust identity preservation while maintaining natural pose diversity. We believe that this benchmark can facilitate collaborative efforts to address the challenging problem of Multi-Human generation.

References

[1] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan's latent space. *ACM Transactions on Graphics (TOG)*, 39(6):1–17, 2020.

[2] George Arfken. *Mathematical Methods for Physicists*. Academic Press, Inc., San Diego, third edition, 1985.

[3] Clément Bonnet, Ariel N Lee, Franck Wertel, Antoine Tamano, Tanguy Cizain, and Pablo Ducru. From text to pose to image: Improving diffusion model control and quality. *arXiv preprint arXiv:2411.12872*, 2024.

[4] Shubhankar Borse, Kartikeya Bhardwaj, Mohammad Reza Karimi Dastjerdi, Hyojin Park, Shreya Kadambi, Shobitha Shivakumar, Prathamesh Mandke, Ankita Nayak, Harris Teague, Munawar Hayat, et al. Subzero: Composing subject, style, and action via zero-shot personalization. *arXiv preprint arXiv:2502.19673*, 2025.

[5] Shubhankar Borse, Farzad Farhadzadeh, Munawar Hayat, and Fatih Porikli. Disco: Reinforcement with diversity constraints for multi-human generation. *arXiv preprint arXiv:2510.01399*, 2025.

[6] Anthony Chen, Jianjin Xu, Wenzhao Zheng, Gaole Dai, Yida Wang, Renrui Zhang, Haofan Wang, and Shanghang Zhang. Training-free regional prompting for diffusion transformers. *arXiv preprint arXiv:2411.02395*, 2024.

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[9] 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*, pages 4690–4699, 2019.

[10] 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.

[11] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36:52132–52152, 2023.

[12] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.

[13] Huiguo He, Qiuyue Wang, Yuan Zhou, Yuxuan Cai, Hongyang Chao, Jian Yin, and Huan Yang. Improving multi-subject consistency in open-domain image generation with isolation and reposition attention. *arXiv preprint arXiv:2411.19261*, 2024.

[14] Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[15] 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.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[17] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3707–3714, 2025.

[18] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

[19] Tero Karras, Timo Aittala, and Samuli Laine. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[21] 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.

[22] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *European Conference on Computer Vision*, pages 253–270. Springer, 2024.

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

[24] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023.

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

[26] Duong H Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. *arXiv preprint arXiv:2411.16318*, 2024.

[27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[28] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*, 2024.

[29] Chaojie Mao, Jingfeng Zhang, Yulin Pan, Zeyinzi Jiang, Zhen Han, Yu Liu, and Jingren Zhou. Ace++: Instruction-based image creation and editing via context-aware content filling. *arXiv preprint arXiv:2501.02487*, 2025.

[30] OpenAI. Gpt-image-1. `https://platform.openai.com/docs/models/gpt-image-1`, 2025.

[31] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang. Lambda-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space. *arXiv preprint arXiv:2402.05195*, 2024.

[32] 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.

[33] PopulationPyramid.net. World population pyramid 2024, 2024.

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[35] Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. *arXiv preprint arXiv:2310.03502*, 2023.

[36] 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)*, pages 10684–10695, 2022.

[37] SG_161222. Realistic vision v1.3 (and other versions), 2023.

[38] Zhicheng Sun, Zhenhao Yang, Yang Jin, Haozhe Chi, Kun Xu, Liwei Chen, Hao Jiang, Yang Song, Kun Gai, and Yadong Mu. Rectifid: Personalizing rectified flow with anchored classifier guidance. *Advances in Neural Information Processing Systems*, 37:96993–97026, 2024.

[39] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024.

[40] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024.

[41] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024.

[42] 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.

[43] 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*, pages 1–20, 2024.

[44] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024.

[45] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[47] Xizhuo Yu, Chaojie Fan, Zhizhong Zhang, Yongbo Wang, Chunyang Chen, Tianjian Yu, and Yong Peng. Identity-aware infrared person image generation and re-identification via controllable diffusion model. *Pattern Recognition*, 165:111561, 2025.

[48] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:11127–11150, 2023.

[49] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims are reflected throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations for every benchmarked approach are provided in Section 5, E

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical result is provided in the paper

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix C And Section. 3 provides in-depth details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and data are provided in the link. Readme/croissant files are added for help in reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix D provides experimental details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars across multiple seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Provided in appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed te code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Societal impacts discussed in Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No models released. The data poses no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We clearly credit every dataset/model.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Data and code are documented in Readme and Croissant format.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowd sourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No study on human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We cite each MLLM used and clearly specify them.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendices

A Contents

This appendix provides supplementary material to accompany the main paper. It includes a detailed breakdown of the data distribution used, further specifics on our proposed methodology, implementation details for baselines and our approach, extended quantitative and qualitative results including ablation studies and failure cases, and a discussion on the societal impact of our work. The following sections detail these aspects:

B Data Distribution

In this Section, we provide a detailed description of the sampled faces. This is supplementing Figure 3 in the main text.

Attribute	Category	Percentage (%)
Age	Young Adult (16-35)	43.17
	Middle Age (36-60)	42.72
	Aged (60+)	14.11
Gender	Male	49.24
	Female	50.76
Status	Anonymous	81.48
	Celebrity	18.52
Data	Real	72.41
	Synthetic	27.59
Ethnicity	White	16.52
	Black	15.75
	South Asian	16.64
	East Asian	16.72
	Hispanic/Latin	16.73
	Middle Eastern/Other European	17.64

Table B.1: Data Distribution by Attribute. All the labels are obtained by Gemini-Flash-2.0.

C Additional Details on Proposed Approach

In this Section, we provide detailed explanation of our proposed Implicit Region Assignment strategy behind the MH-Omnigen and MH-IR-Diffusion methods. This is an extension of Section 3.

C.1 Implicit Region Assignment

To construct the attention mask $\mathbf{A_{iso}}$ described in Section 3, we need a region of interest (ROI) set $\mathcal{R}_k \subseteq \mathcal{D}_{\text{latent}}$ for every reference image I_k. To facilitate the generation of multi-human images in the wild, we propose an implicit region assignment strategy that utilizes intermediate attention scores and Hungarian matching to assign each reference image to a selected region-of-interest.

MH-Omnigen: For unified architectures, we propose a two-stage process to determine ROIs implicitly from the model's own understanding and the intermediate latent representation $\mathbf{O}_{int,t}$. This process involves first identifying areas in the latent space which have a high self-attention overlap with each reference image. Next, we segment the estimated image from an intermediate timestep in and assign these segments to the reference images using hungarian matching.

Algorithm 1 Find attention-based similarity maps for Reference Images

Inputs:

- Unified Multimodal Diffusion Model U.
- Intermediate latent $\mathbf{O}_{int,t} \in \mathbb{R}^{D\times D}$ at timestep t.
- Set of K input reference images: $\{I_1, \ldots, I_K\}$.
- Image token index sets: $\mathcal{D}_{\text{img},k} \subseteq \{1, \ldots, L\}$ for $k = 1, \ldots, K$.
- Latent token index set: $\mathcal{D}_{\text{latent}} \subseteq \{1, \ldots, L\}$.
- Reshape map $\mathcal{M}_{\text{reshape}} : \mathbb{R}^{|\mathcal{D}_{\text{latent}}|} \to \mathbb{R}^{D\times D}$, where $|\mathcal{D}_{\text{latent}}| = D \times D$.
- Set of H layers to probe: $\mathcal{L} = \{l_1, \ldots, l_H\}$.

Outputs:

- Set of K similarity maps for every reference image: $\{\mathbf{S}_k \mid k = 1, \ldots, K\}$, each $\mathbf{S}_k \in \mathbb{R}^{D\times D}$.

▷ Initialize similarity maps
for $k = 1, \ldots, K$ **do**
  Initialize $\mathbf{S}_k$ as a $D \times D$ zero matrix.
  $\mathbf{S}_k[p, q] = 0$ for all (p, q).
end for

▷ Iterate through layers
for l in $\mathcal{L}$ **do**
  ▷ Get attention map for layer l at timestep t
  Let $\mathbf{P}^{(l)} \in \mathbb{R}^{L\times L}$ be the self-attention map from layer l.
  $\mathbf{P}^{(l)}_{ij}$ is attention from query token i to key/value token j, for $i, j \in \{1, \ldots, L\}$.
  ▷ Iterate through all reference images I_k
  for $k = 1, \ldots, K$ **do**
    ▷ Slice attention maps for latent and aggregate over tokens corresponding to I_k
    Define flat vector $\mathbf{V}_{k,l} \in \mathbb{R}^{|\mathcal{D}_{\text{latent}}|}$.
    For each $i \in \mathcal{D}_{\text{latent}}$, $\mathbf{V}_{k,l}[i] = \sum_{j\in\mathcal{D}_{\text{img},k}} \mathbf{P}^{(l)}_{ij}$.
    ▷ Reshape $\mathbf{V}_{k,l}$ into spatial map $\mathbf{SM}_{k,l} \in \mathbb{R}^{D\times D}$
    $\mathbf{SM}_{k,l} = \mathcal{M}_{\text{reshape}}(\mathbf{V}_{k,l})$.
    ▷ Accumulate $\mathbf{SM}_{k,l}$ into $\mathbf{S}_k$
    $\mathbf{S}_k \leftarrow \mathbf{S}_k + \mathbf{SM}_{k,l}$ ▷ Element-wise addition
  end for
end for

return $\{\mathbf{S}_k \mid k = 1, \ldots, K\}$

The first stage, detailed in Algorithm 1, computes attention-based similarity maps $\{\mathbf{S}_k\}_{k=1}^{K}$. For each reference image I_k, this algorithm probes the layers $\mathcal{L}$ of the unified multimodal diffusion model U. It aggregates the attention from latent tokens $i \in \mathcal{D}_{\text{latent}}$ to the image tokens $j \in \mathcal{D}_{\text{img},k}$ corresponding to I_k. This aggregated attention is reshaped via $\mathcal{M}_{\text{reshape}}$ into a $D \times D$ spatial map $\mathbf{SM}_{k,l}$ for each layer l. These layer-specific maps are then accumulated to form the final similarity map $\mathbf{S}_k \in \mathbb{R}^{D\times D}$. Each $\mathbf{S}_k$ thus highlights regions in the $D \times D$ latent space which exhibit strong attention probability with the k-th reference image.

The second stage, outlined in Algorithm 2, uses the similarity maps $\mathbf{S}_k$ (generated by Algorithm 1) to derive the final binary segmentation maps $\mathcal{R}_k \in \{0,1\}^{D\times D}$ which represent the ROIs. This algorithm takes as input the predicted latent at $t = 0$, $\hat{\mathbf{O}}_{\text{int},0}$ and outputs a set of K assigned binary segmentation maps $\{\mathcal{R}_k \in \{0,1\}^{D\times D}\}_{k=1}^{K}$.

We first generate the estimated image from the predicted latent $\hat{\mathbf{O}}_{\text{int},0}$ by passing through the VAE Decoder D_{VAE}. This is then segmented using the Segment Anything 2 [34] model, denoted SAM, to produce initial segmentation masks $\{\mathbf{M}_{seg,j}\}$. These masks are subsequently refined by Non-Maximum Suppression (NMS) to yield Q binary segmentation maps $\{\mathbf{G}_q\}_{q=1}^{Q}$. These maps $\mathbf{G}_q$ are transformed to $D \times D$ resolution required for comparison with the similarity maps $\mathbf{S}_k$, and represent the candidate regions within the decoded image. If no regions survive NMS ($Q = 0$), the algorithm initializes all output maps $\mathcal{R}_k$ to one and terminates.

Algorithm 2 Assign Regions for each Reference Image

Inputs:

- Predicted latent at $t = 0$: $\hat{\mathbf{O}}_{\text{int},0} \in \mathbb{R}^{dim_{lat}}$.
- VAE Decoder D_{VAE}.
- Segmentation Model SAM.
- NMS threshold θ_{NMS}.
- Set of K similarity maps $\{\mathbf{S}_k \in \mathbb{R}^{D\times D}\}_{k=1}^{K}$ (from Alg. 1).
- Number of input reference images K.

Outputs:

- Set of K assigned generated segmentation maps $\{\mathcal{R}_k \in \{0,1\}^{D\times D}\}_{k=1}^{K}$.

▷ Step 1: Generate Segmentation Maps from $\hat{\mathbf{O}}_{\text{int},0}$
$\{\mathbf{M}_{seg,j}\}_{j=1}^{Q'} \leftarrow SAM(D_{VAE}(\hat{\mathbf{O}}_{\text{int},0}))$. ▷ Segment VAE decoder output.
$\{\mathbf{G}_q \in \{0,1\}^{D\times D}\}_{q=1}^{Q} \leftarrow NMS(\{\mathbf{M}_{seg,j}\}_{j=1}^{Q'}, \theta_{NMS})$. ▷ Q maps post-NMS, at $D \times D$ res.
▷ Step 2: Construct Cost Matrix $\mathbf{C} \in \mathbb{R}^{K\times Q}$
for $k = 1, \ldots, K$ **do**
    for $q = 1, \ldots, Q$ **do**
        $\mathbf{C}[k, q] = -\sum_{p=1}^{D}\sum_{r=1}^{D}(\mathbf{S}_k[p, r] \cdot \mathbf{G}_q[p, r])$. ▷ Negative overlap score.
    end for
end for

▷ Step 3: Apply Hungarian Algorithm
$\mathcal{A} \leftarrow Hungarian(\mathbf{C})$. ▷ $\mathcal{A}$ is set of optimal (k, q) pairs.

▷ Step 4: Formulate Final Output $\mathcal{R}_k$
Initialize $\mathcal{R}_k$ as $D \times D$ ones matrix for $k = 1, \ldots, K$.
for each assignment $(k^*, q^*) \in \mathcal{A}$ **do**
    $\mathcal{R}_{k^*} \leftarrow \mathbf{G}_{q^*}$.
end for

return $\{\mathcal{R}_k\}_{k=1}^{K}$.

If $Q > 0$, a $K \times Q$ cost matrix $\mathbf{C}$ is constructed to evaluate the compatibility between each reference image I_k (via its similarity map $\mathbf{S}_k$) and each generated segmentation map $\mathbf{G}_q$. We compute the

cost of assigning generated region $\mathbf{G}_q$ to reference image I_k, $\mathbf{C}[k, q]$, from the overlap between $\mathbf{G}_q$ and the similarity map $\mathbf{S}_k$ (Algorithm 2, Line 14). The Hungarian algorithm [23] is then applied to minimize the cost matrix $\mathbf{C}$. This yields a set $\mathcal{A}$ of optimal assignment pairs (k, q), where each reference image k is matched to at most one generated map q.

Finally, the output maps $\{\mathcal{R}_k\}$ are first initialized to $D \times D$ ones matrices. For each optimal assignment $(k^*, q^*) \in \mathcal{A}$, the corresponding generated segmentation map $\mathbf{G}_{q^*}$ is assigned as the output map $\mathcal{R}_{k^*}$ for the reference image I_{k^*}. If a reference image I_k is not part of any assignment in $\mathcal{A}$ (e.g., if $K > Q$), its $\mathcal{R}_k$ remains a ones map.

Once we find the region assignments $\{\mathcal{R}_k\}_{k=1}^{K}$, we apply the Unified Isolated Attention approach explained in Section 3 to generate the final image.

MH-IR-Diffusion: For models like IR-Diffusion that involve generating an initial image I_{gen}, followed by a final image with specific identities, we can determine regions $\mathcal{R}_k$ by directly matching facial identity cues. This method uses ArcFace embeddings to associate faces segmented from I_{gen} with the K input reference images I_k. The segmentation mask of an assigned face in I_{gen} serves as the ROI $\mathcal{R}_k$.

We show our approach in Algorithm 3. Initially, SAM and NMS are used to obtain Q distinct face segmentation masks $\{\mathbf{G}_q\}$ from I_{gen}. ArcFace embeddings are then computed for these generated face regions $\{\mathbf{e}_{gen,q}\}$ and for the reference images $\{\mathbf{e}_{ref,k}\}$. A cost matrix $\mathbf{C}$ is built using the cosine dissimilarity ($1 -$ cosine_similarity) between these embeddings. The Hungarian algorithm then finds the optimal assignment $\mathcal{A}$ between reference images and generated faces. For each match, the corresponding mask $\mathbf{G}_{q^*}$ is designated as $\mathcal{R}_{k^*}$. These pixel-space masks $\mathcal{R}_k$ are then available for subsequent processing steps.

D Baselines and Implementation Details

In this Section, we highlight our Implementation Details for our baselines. Every experiment was performed on an Nvidia Tesla A100 GPU.

OmniGen. We used the official implementation of OmniGen[2] and prompted it for multi-human with and without pose conditioning. We run the default settings of 50-step inference, with a text-guidance of 2.5. For Task-1, we set image guidance scale at 2.0, and for Task-2, we found the best results with 2.8. We implemented MH-OmniGen over this repository.

UniPortrait. We used the official implementation of UniPortrait[3] and used all the default hyper-parameters. From the original settings, we perform a 25-step inference at a guidance scale of 7.5. For pose conditioning, UniPortrait adopts controlnet, for which we set the guidance scale at 1.

FastComposer. We used the official implementation of FastComposer[4] and used all the default hyper-parameters. From the original settings, we perform a 50-step inference at a guidance scale of 5.

OMG. We used the official implementation of OMG[5], and used the the InstantID version. As the performance was poor without ControlNet, we only report the performance with regional priors, and used all the default hyper-parameters. The inference is run in 50 steps, using a CFG scale of 3.0. The controlnet and InstantID models both were weighted at 0.8, as per original implementation. To make the performance suitable to MultiHuman, we modified their detection algorithm to match all detected humans (instead of the deafult matching with "man" and "woman").

Regional-PuLID. We used the official implementation of Regional-PuLID[6] and made it compatible with Multi-human generation with default box priors based on number of humans. We found best results with base ratio set to 0.3. We kept the remaining hyperparameters at default settings. These include 24 inference steps and a guidance scale of 3.5.

[2] https://github.com/VectorSpaceLab/OmniGen

[3] https://github.com/junjiehe96/UniPortrait/tree/main

[4] https://github.com/mit-han-lab/fastcomposer/tree/main

[5] https://github.com/kongzhecn/OMG.git

[6] https://github.com/instantX-research/Regional-Prompting-FLUX

Algorithm 3 Assign Regions for MH-IR-Diffusion using ArcFace Embeddings

Inputs:

- Generated image I_{gen} ($H \times W$).
- Set of K generated reference images $\{I_k\}_{k=1}^{K}$.
- Segmentation Model SAM (for face segmentation).
- NMS threshold θ_{NMS}.
- ArcFace embedding function $ArcFace$.

Outputs:

- Set of K assigned face segmentation masks $\{\mathcal{R}_k \in \{0,1\}^{H \times W}\}_{k=1}^{K}$.

▷ Step 1: Segment faces in I_{gen} and compute their ArcFace embeddings
$\{\mathbf{M}_{seg,j}\}_{j=1}^{Q'} \leftarrow SAM(I_{gen})$.
$\{\mathbf{G}_q \in \{0,1\}^{H \times W}\}_{q=1}^{Q} \leftarrow NMS(\{\mathbf{M}_{seg,j}\}_{j=1}^{Q'}, \theta_{NMS})$.
for $q = 1, \ldots, Q$ **do**
    $\mathbf{e}_{gen,q} \leftarrow ArcFace(I_{gen}, \mathbf{G}_q)$. ▷ Embedding for face in I_{gen} at mask $\mathbf{G}_q$.
end for

▷ Step 2: Compute ArcFace embeddings for I_k
for $k = 1, \ldots, K$ **do**
    $\mathbf{e}_{ref,k} \leftarrow ArcFace(I_k)$.
end for

▷ Step 3: Construct Cost Matrix $\mathbf{C} \in \mathbb{R}^{K \times Q}$
for $k = 1, \ldots, K$ **do**
    for $q = 1, \ldots, Q$ **do**
        $\mathbf{C}[k, q] = 1 - \text{cosine\_similarity}(\mathbf{e}_{ref,k}, \mathbf{e}_{gen,q})$.
    end for
end for

▷ Step 4: Apply Hungarian Algorithm
$\mathcal{A} \leftarrow Hungarian(\mathbf{C})$.

▷ Step 5: Formulate Final Output $\mathcal{R}_k$
Initialize $\mathcal{R}_k$ as $H \times W$ zeros matrix for $k = 1, \ldots, K$.
for each assignment $(k^*, q^*) \in \mathcal{A}$ **do**
    $\mathcal{R}_{k^*} \leftarrow \mathbf{G}_{q^*}$.
end for

return $\{\mathcal{R}_k\}_{k=1}^{K}$.

LoRA/MuDI. We used the official implementation of MuDI[7] and trained a single LoRA for every sample. We used the default settings for training, with 2000 steps at a learning rate of $1e^{-4}$. We train in two settings, with a single view per face and 5 views per face. During in inference, we kept the default setting with an inference of 50 steps and guidance scale of 5. We run inference with both LoRA and MuDI using the provided examples. For LoRA with pose, we use the SDXL openpose controlnet with a scale of 1.0.

MIP-Adapter. We used the official implementation of MIP-Adapter[8], which builds upon a pretrained IP-Adapter and SDXL model, and incorporates OpenCLIP-ViT-bigG/14 as the image encoder. We loaded the released MIP-Adapter weights into this framework to better support multi-subject generation. For multi-human generation, we used prompts from our MultiHuman-Testbench and

[7] `https://github.com/agwmon/MuDI`
[8] `https://github.com/hqhQAQ/MIP-Adapter`

adopted the DDIM sampler with 30 inference steps and a guidance scale of 7.5. The IP-Adapter scale was set to 0.75. For pose-based regional conditioning, we followed the ControlNet implementation[9].

IP-Adapter. While the pretrained IP-Adapter provides strong performance, it is not designed to directly support multiple reference images. To address this, we adopted the MIP-Adapter framework to leverage its mechanism of weight merging within the adapter layers, without loading the MIP-Adapter weights. Instead, we retained only the pretrained IP-Adapter weights in our implementation. Sampling parameters, including 30 inference steps, guidance scale of 7.5, and IP-Adapter scale of 0.75, were set identically to those in MIP-Adapter for consistency.

RectifID. We utilize the official implementation provided by the authors[10]. Following their setup, we adopt a modified version of Stable Diffusion 1.5 released by Perflow[11], which is pretrained on the LAION-Aesthetic-5+ dataset with a particular focus on human faces and subject-centric generation. For the inversion process, we perform 50 sampling steps using classifier-free guidance with a guidance scale of 3.0. In experiments involving ControlNet, all settings are kept identical except for the use of ControlNet weights[12].

λ-Eclipse. We employed the official implementation provided by the authors[13], a model designed for multi-concept personalized text-to-image generation. This model operates within the CLIP latent space and is specifically tailored to work with the Kandinsky v2.2 [35] diffusion image generator. For inference, we employed the DDIM sampler with 50 steps and a guidance scale of 7.5. All experiments were conducted using the provided scripts and configurations to ensure consistency and reproducibility.

E Additional Quantitative Results

In this Section, we provide additional Quantitative analysis on the MultiHuman Testbench, as an extension to Section 5 of the paper. The main results from Section 5 are summarized in the Radar graphs in Figure E.1.

E.1 Performance across varying number of people

In this Section, we study Multi-ID similarity scores and Person Count for methods performing Task 1, across different number of people. This is an extension to Table 3in the main paper.

Summarized in Table E.1, we consistently observe that the ability to maintain ID Similarity deteriorates across all models as the number of faces in the generated image increases from two to five. This indicates a general bias where identity preservation becomes significantly more challenging with growing scene complexity. However, the magnitude and nature of this drop vary between methods. Methods such as UniPortrait, OmniGen, and MH-OmniGen start with relatively high ID Similarity scores for one or two people, indicating strong initial performance. As the person count increases, they experience substantial absolute drops in ID Similarity. For instance, MH-OmniGen's score falls from 65.3 at two people to 38.0 at five people. Fastcomposer begins with a significantly lower ID Similarity even for just two people (15.3) and suffers the most dramatic percentage drop, falling to 5.9 at five people. GPT-Image-1 starts at a moderate ID Similarity score for two people (31.8) and exhibits the least severe relative decrease in performance as the person count increases, resulting in a score of 24.9 at five people. From the Person Count results, it is clear that GPT-Image-1 is more accurate in generating images with the correct number of people, compared to other methods, which fail often for more than three people.

E.2 Measuring Bias in Multi-Human Generation

Using the labeled attributes provided with the data, we provide a study on the biases for each method in Task 1 and Task 2. We measure the single-person ID similarity and report it across various splits: Ethnicity(E.2), age, gender and status(E.3). Based on the ID-Similarity scores highlighted in red in

[9] https://huggingface.co/thibaud/controlnet-openpose-sdxl-1.0

[10] https://github.com/feifeiobama/RectifID

[11] https://huggingface.co/hansyan/perflow-sd15-dreamshaper

[12] https://huggingface.co/lllyasviel/control_v11p_sd15_openpose

[13] https://github.com/eclipse-t2i/lambda-eclipse-inference

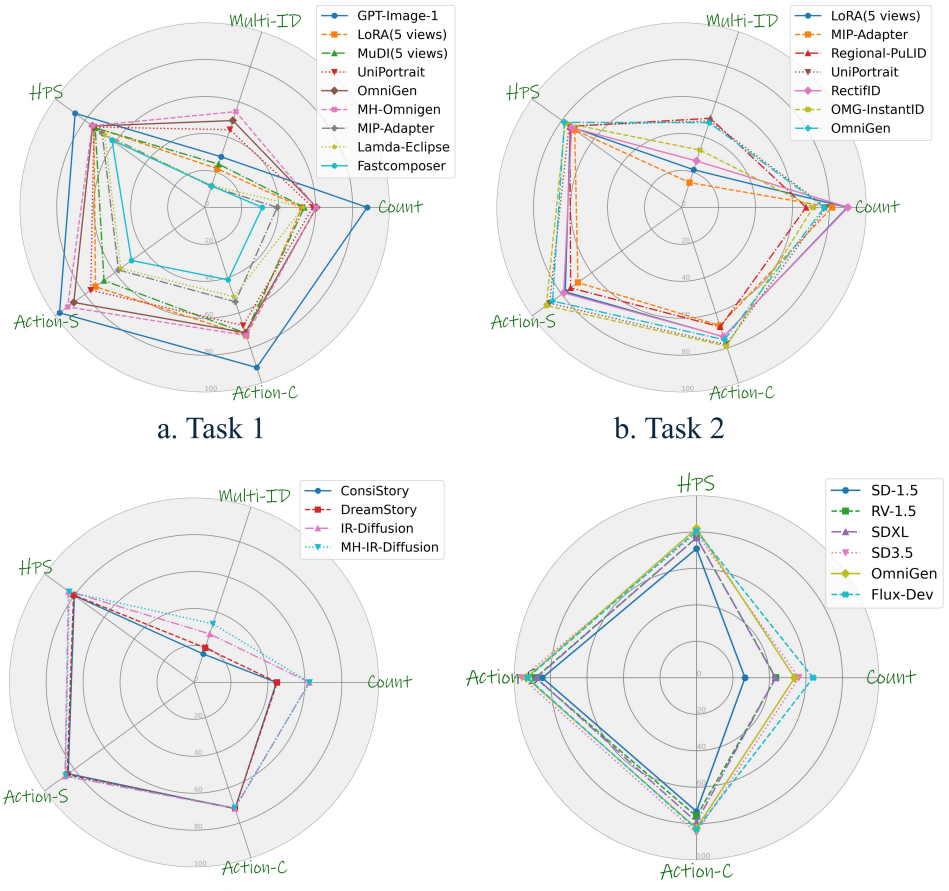


Figure E.1: **Radar Graphs**. Visualizing performance on Tasks 1,2,3,4 using Radar Graphs.

Model	Multi-ID Similarity					Person Count				
	2	3	4	5	Avg(1-5)	2	3	4	5	Avg(1-5)
Task 1: Reference-based Multi-Human Generation in the Wild										
GPT-Image-1	31.8	29.5	27.8	24.9	28.8	90.7	91.8	89.5	75.3	87.9
Fastcomposer	15.3	7.4	7.2	5.9	12.2	62.9	11.2	3.2	1.1	31.2
Uniportrait	56.5	46.4	33.8	28.6	44.2	90.6	76.3	23.7	14.1	58.5
OmniGen	60.8	52.3	42.2	35.2	49.4	88.8	88.0	23.2	21.6	60.5
MH-OmniGen	65.3	60.4	45.1	38.0	54.5	91.2	87.5	22.4	19.7	60.3

Table E.1: Studying the ID similarity and Person count metrics for different number of people, for Multi-Human Generation in the Wild.

Table E.2 and Table E.3, biases are evident in how different models and their backbones perform. These biases are indicated by deviations from the average score for each row. A darker red shades signifies higher bias.

In Task 1, GPT-Image-1 (GPT-4o) shows a positive bias for South-Asian identities and a positive bias for Aged individuals. UniPortrait (RV1.5) exhibits positive biases favoring White and East-Asian faces while underperforming on Black individuals. Additionally, it heavily underperforms on Aged individuals by demographic type, offset by a positive bias for Young Adults. Fastcomposer (SD1.5) shows minimal racial/ethnic bias but has a light negative bias for Celeb faces. OmniGen (Phi-3) and MH-OmniGen (Phi-3) generally display less pronounced biases in Task 1, showing mostly light biases favoring White and South-Asian faces by race/ethnicity, suggesting more balanced performance compared to the rest.

Model	Backbone	ID-Similarity					
		White	Black	South-Asian	East-Asian	Hispanic	Middle-East
Task 1: Reference-based Multi-Human Generation in the Wild							
GPT-Image-1	GPT-4o	26.4	28.8	30.2	26.2	27.5	28.2
UniPortrait	RV1.5	41.1	36.8	38.3	41.8	37.5	38.4
Fastcomposer	SD1.5	8.7	8.8	9.3	9.5	9.2	8.9
OmniGen	Phi-3	44.0	45.1	47.0	45.0	45.5	44.3
MH-OmniGen	Phi-3	48.4	49.3	51.6	50.0	49.7	49.0
Task 2: Reference-based Multi-Human Generation with Regional Priors							
UniPortrait	RV1.5	48.9	43.6	45.3	47.3	43.3	45.3
RectifID	SD1.5	17.6	16.9	19.2	21.1	16.3	17.1
Regional-PuLID	Flux	47.4	44.1	47.9	52.6	47.3	46.0
OMG-InstantID	SDXL	26.8	24.3	27.0	30.6	25.9	25.1
OmniGen	Phi-3	42.7	42.6	44.9	43.6	41.4	41.2

Table E.2: Multi-Human Generation with Reference Images: Multi-Human Tuning-Free Models with ID-Similarity Metrics by Race/Ethnicity (Backbone Removed)

Turning to Task 2, where regional priors are used, the patterns of bias shift for some models. UniPortrait (RV1.5) continues to show biases favoring White and East-Asian faces and against Black and Hispanic faces, favoring female faces to male faces, and heavily favoring young adults. RectifID (SD1.5) shows a bias favoring East-Asian faces by race/ethnicity. Regional-PuLID (Flux) displays significant biases, with a strong positive bias for East-Asian individuals and a negative bias against Black faces by ethnicity. By demographic, Regional-PuLID exhibits strong biases against Males and Aged faces, while strongly favoring Female, Young Adult, and Celeb identities. OMG-InstantID (SDXL) shows a bias against Black faces and favoring East-Asian faces, and favors Young Adults. OmniGen (Phi-3) in Task 2 shows less prominent biases, with a bias favoring South-Asian faces and light biases against Hispanic and Middle-East faces by ethnicity.

Model	Backbone	ID-Similarity						
		Male	Female	Young Adult	Middle Aged	Aged	Celeb	Anonymous
Task 1: Reference-based Multi-Human Generation in the Wild								
GPT-Image-1	GPT-4o	29.8	26.0	26.0	28.8	30.8	28.1	26.6
UniPortrait	RV1.5	37.5	40.5	40.7	37.7	30.7	37.6	39.3
Fastcomposer	SD1.5	9.0	9.1	9.2	9.1	8.5	7.1	9.5
OmniGen	Phi-3	44.9	45.4	44.3	46.0	45.4	45.5	43.7
MH-OmniGen	Phi-3	49.5	49.8	49.0	49.9	50.7	48.7	49.9
Task 2: Reference-based Multi-Human Generation with Regional Priors								
UniPortrait	RV1.5	43.5	47.7	48.2	44.0	42.9	46.9	45.3
RectifID	SD1.5	17.4	18.6	19.1	17.4	16.1	19.5	17.7
Regional-PuLID	Flux	42.3	52.5	51.7	45.4	41.5	50.5	46.9
OMG-InstantID	SDXL	25.0	28.2	28.7	25.1	24.9	24.9	26.8
OmniGen	Phi-3	43.6	42.1	41.8	43.2	44.6	42.6	42.9

Table E.3: Multi-Human Generation with Reference Images: Multi-Human Tuning-Free Models with ID-Similarity Metrics by Demographic and Type (Backbone Removed)

E.3 Effect of Our Pose Priors

Table E.4 shows the effect of adding our human-rectified regional pose priors. As observed, all metrics significantly improve in case of most methods. There is a slight drop in Multi-ID and Action-S for OmniGen, and a slight drop in Multi-ID in LoRA. Overall, regional pose priors can help generate significantly better results, as they make the task easier. However, this benefit comes with a severe hit to usability, which is why solving Task 1 is important.

F Additional Qualitative Results

In this Section, we provide additional Qualitative results to supplement the ones in Section 5 of the main text.

Model	Count		Multi-ID		HPS		Action-S		Action-C	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
LoRA(5 views)	52.6	**89.6**	**22.0**	21.4	25.9	**26.0**	73.0	**77.7**	72.9	**73.6**
MIP-Adapter	39.2	**81.5**	11.9	**14.1**	24.0	**25.0**	57.6	**69.2**	53.7	**67.2**
UniPortrait	58.5	**78.3**	44.2	**49.2**	25.9	**26.3**	76.2	**88.2**	67.2	**78.1**
RectifID	37.8	**90.1**	18.6	**26.4**	24.8	**25.7**	67.3	**78.7**	68.2	**73.5**
OmniGen	60.5	**77.2**	**49.4**	48.2	26.2	**27.4**	**87.5**	86.2	71.3	**75.3**

Table E.4: Comparison of Multi-Human Generation Metrics With and Without Regional Priors for Tuning-Free Models

F.1 Improvements from MH-OmniGen

Figure F.1 shows more qualitative results on Multi-Human generation in the wild, using OmniGen and MH-OmniGen. As observed, MH-OmniGen is able to correct many instances of subject blending using our proposed Unified Region Isolation and Implicit Matching algorithm.

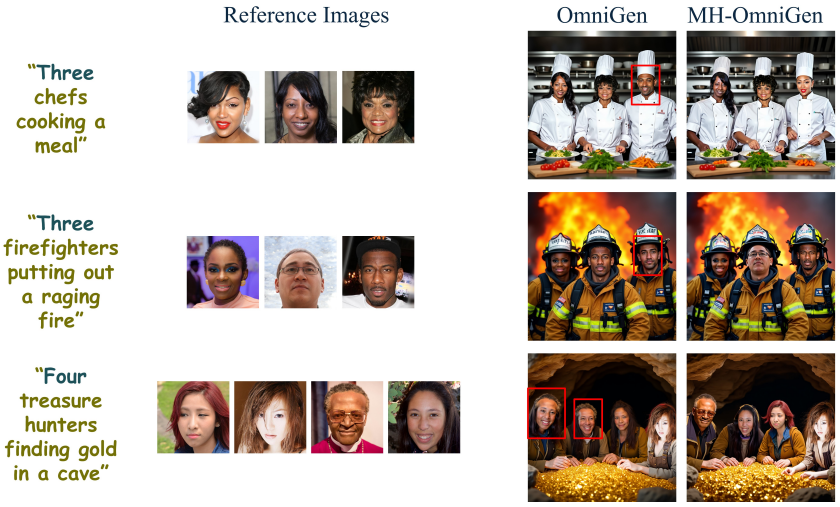


Figure F.1: **Qualitative Results for MH-OmniGen v/s OmniGen on Task 1**. As observed, MH-OmniGen is able to significantly improve OmniGen by reducing ID leakage, which is highlighted with red boxes.

F.2 Qualitative results for Task 2: Regional Priors

Figure F.2 shows the best performing methods: RectifID, UniPortrait, LoRA, OmniGen and Regional-PuLID(with box priors) on Task 2: Reference-based Multi-Human Generation with Regional Priors. As observed in the figure, OmniGen, UniPortrait and Regional-PuLID show best results. However, it is clear that each method has severe limitations including incorrect count accuracy and underperformance on ID similarity. This underlines a huge scope for improvement.

F.3 Qualitative results for Task 4: Text-to-Image generation

Figure F.3 shows the best performing methods on Task 4: Text-to-Image Generation for Multiple Humans (with no reference images). The methods on display are RV1.5, SDXL, SD3.5, OmniGen and Flux. Flux and OmniGen show best results. However, it is clear that all methods show limitations in terms of count accuracy. Additionally, every method is susceptible to generating faces of people with similar attributes (age, race, gender). We believe that there is a heavy scope for improvement for Generation Models in this regard.

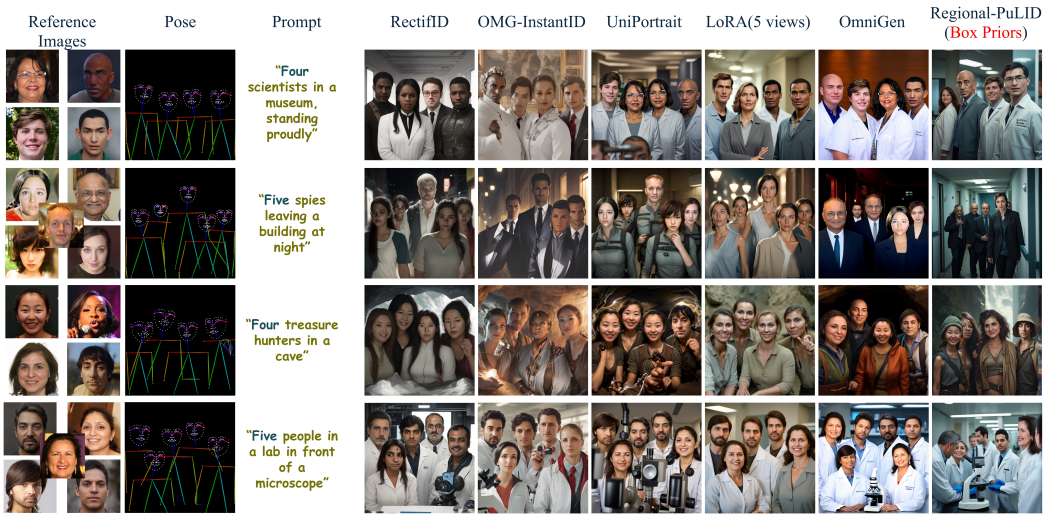


Figure F.2: **Qualitative Results on Multi-Human Generation with Pose conditioning**. The image shows the best performing methods: RectifID, UniPortrait, LoRA, OmniGen and Regional-PuLID(with box priors). OmniGen, UniPortrait and Regional-PuLID show best results albeit with severe limitations.

G Societal Impact

With MultiHuman-Testbench, we aim to make significant advancements in AI-driven multi-human image generation, and we anticipate substantial positive societal benefits. By encouraging the development of models which accurately depict diverse individuals across age, ethnicity, and gender while preserving their identities in complex scenes, we hope to contribute to more equitable and inclusive digital media. We envision that our benchmark can enhance creative tools for artists and developers, enrich AR/VR/XR experiences, and improve assistive technologies. Furthermore, we believe that our proposed standardized evaluation suite will accelerate research and offer clearer insights into generation model capabilities.

However, we also recognize that this progress amplifies societal risks. The capability for highly realistic multi-human image generation increases the potential for deepfakes which could be used in misinformation campaigns or impersonation, thereby posing threats to individual privacy and societal trust. Finally, we acknowledge that the increasing sophistication of these generative tools raises concerns about job displacement in creative sectors. Hence, we request the broader community to proactively engage in developing ethical frameworks, and responsible use guidelines.

Figure F.3: **Qualitative Results on Multi-Human Generation for Text-to-Image models**. The image shows the best performing methods on Task 4: RV1.5, SDXL, SD3.5, OmniGen and Flux. Flux and OmniGen show best results albeit all methods show limitations in terms of count accuracy.