MULTIREF: Controllable Image Generation with Multiple Visual References

Anonymous CVPR submission

Paper ID XXX

Abstract

Visual designers naturally draw inspiration from multiple visual references, combining diverse elements and aesthetic principles to create artwork. However, current image generative frameworks predominantly rely on single-source inputs — either text prompts or individual reference images. In this paper, we present a new task called MULTIREF, *which focuses on controllable image generation using multiple visual references. To support this task, we further introduce* MULTIREF-BENCH, *a rigorous evaluation framework comprising 990 synthetic and 1,000 real-world generation samples that require incorporating visual content from multiple reference images. The synthetic samples are synthetically generated through our data engine, with 10 reference types and 32 reference combinations. For assessment, we integrate both rule-based metrics and a fine-tuned MLLM-as-a-Judge model into* MULTIREF-BENCH. *Our experiments across three interleaved image-text models (*i.e.*, OmniGen, ACE, and Show-o) and six agentic frameworks (*e.g.*, ChatDiT and LLM + SD) reveal that even state-of-the-art systems struggle with multi-reference conditioning, with the best model OmniGen achieving only 66.6% in synthetic samples and 79.0% in real-world cases on average comparing to golden answer. These findings provide valuable directions for developing more flexible and human-like creative tools that can effectively integrate multiple sources of visual inspiration.*

1. Introduction

Digital artists and visual designers often create a new scene by blending elements from multiple source images: a color palette from a *Monet* painting, the architectural form of the *Eiffel Tower* from a photograph, and the texture from a *hand-drawn sketch*. Artists draw inspiration from multiple visual references, mixing diverse elements. This multi-reference creative process allows far more controllable image creation than relying on a single source of inspiration (Figure 1). However, current tools for this artistic process remain too primitive to be directly useful.

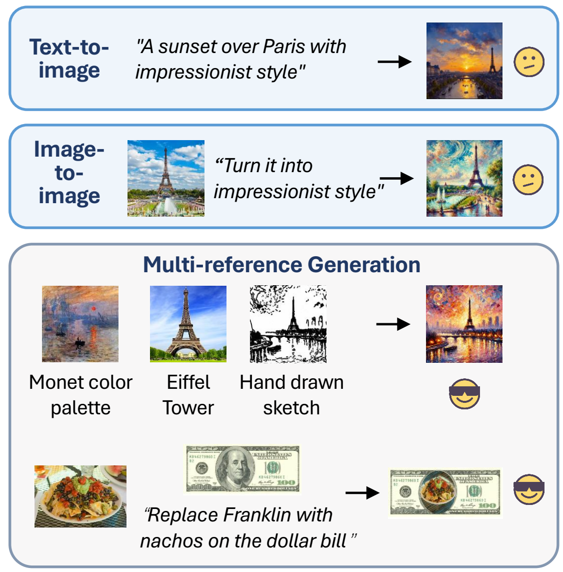

Figure 1. Image generation conditioned on multiple visual references provide more controllable and creative digital art generation than single image or textual reference.

However, today's image generators predominantly rely on single-source conditioning—either a text prompt (*i.e.*, text-to-image [11, 44]) or one reference image (*i.e.*, image editing [30, 45], image translation [19, 52]) at a time. In essence, asking a modern image generative model to *"paint a scene in the style of Van Gogh with the composition of a photograph"* requires specific prompt engineering [17, 26] or sequential editing [25, 53]. Moreover, visual references may have inconsistent viewpoints, styles, or semantics, and merging them can produce contradictions (*e.g.*, blending a daytime landscape with a night-time style reference). Existing approaches like ControlNet [61] excel at following one conditioning signal (*i.e.*, edge map and depth), but they are not inherently designed to handle *multiple* different conditions at once. Additionally, naively adding more control inputs usually confuses the model, leading to jumbled or degraded outputs [62].

There is a growing need to benchmark current multi-reference generation models. From our investigation, most popular benchmarks in generative modeling focus on text-to-image alignment or single-image editing. For example, IDEA-Bench [32] targets professional design scenarios but still typically deals with one reference at a time or sequential editing. Similarly, ACE [18] evaluates alignment with instructions but does not stress-test combining several images. No established benchmark yet examines models on truly multi-reference tasks for their integrating complexity, making it hard to quantify current research progress.

In this paper, we introduce MULTIREF-BENCH, a benchmark that rigorously evaluates multi-reference generation models with 1,000 *real-world* samples and 990 *synthetic* samples which are programmatically generated. Specifically, we compile challenging user requests from Reddit [50], where both references and ground truth images are real, to evaluate the image generalization ability of models from multiple visual references. Our benchmark encompasses a spectrum of tasks, ranging from relatively straightforward scenarios—such as applying two independent style references—to complex scenarios requiring simultaneous spatial and semantic alignment across multiple sources.

To address the scarcity of multi-reference image generation datasets, we develop a novel synthetic data engine, termed REFBLEND, that efficiently creates diverse training samples. REFBLEND first extracts various visual references (*e.g.*, depth maps, edge drawings, subject masks) from existing images using *state-of-the-art* extraction models. These references are then organized into a compatibility graph structure, where nodes represent individual references and edges indicate which references can be meaningfully combined without contradictions, enabling diverse and high-quality multi-reference to image samples at scale. This engine can readily generate synthetic samples by flexibly combining diverse reference modalities—*e.g.*, a segmentation mask, human sketch, and text caption, each describing different aspects of the intended output—while treating the original image as the corresponding target. By controlling the data generation process, we automatically obtain rich ground-truth pairings of inputs and outputs. Finally, MULTIREF-BENCH contains 100,728 synthetic samples covering 10 reference types and 32 reference combination, *far surpassing* any existing collection in both scale and complexity.

We propose new protocols to evaluate the generations using our benchmark. We leverage rule-based (*e.g.*, MSE for depth) and model-based (*e.g.*, ClipScore [20] for aesthetic) assessments for conditions that require precise evaluation (*e.g.*, depth, mask and bbox) and fine-tuned MLLM-as-a-Judge [5] for semantic-level assessments (*e.g.*, caption, sketch and semantic map) in both reference-following and overall quality with human-annotated scores.

We evaluate three interleaved image-text generation models (*e.g.*, OmniGen [55], ACE [18], Show-o [56]) and 6 agentic frameworks (*e.g.*, ChatDiT [25], LLM [2, 15] + Diffusion [11, 44]). Experimental results reveal that even the most advanced *"general-purpose"* image generators today struggle with multi-reference conditioning. *State-of-the-art* diffusion and autoregressive models that claim to support arbitrary conditioning (*e.g.*, recent unified models) often falter when actually confronted with multiple visual inputs. For instance, a model might capture the style of one reference image well but completely ignore the content from another subject reference. Quantitatively, we observe substantial performance gaps: the best existing model OmniGen achieves only 0.496 of the desired alignment score on multi-reference tasks, compared to its near-perfect performance on single-reference inputs. These results expose a clear weakness in current systems – despite their advertised flexibility, they are not truly equipped for multi-reference generation. By highlighting these shortcomings, our study provides valuable insights and direction for future research.

2. Related Work

Controllable Image Generation. The emergence of controllable image generation has revolutionized artificial intelligence by enabling users to create images that precisely match their specified criteria, including composition [31, 59, 63], style [1, 52], and content elements [6, 7]. ControlNet [61] advanced this field by introducing spatially localized input conditions to pre-trained text-to-image diffusion models through efficient fine-tuning methods. Subsequent research [10, 29, 35, 36] has further enhanced image controllability by implementing additional customization layers and adaptive mechanisms, enabling more sophisticated and precise image generation processes.

Building upon these advancements, some work has studied universal guidance for image generation with diffusion models [3, 34, 37, 40, 41, 57, 62]. While early approaches often required complex, condition-specific adapters, a new generation of unified models has expanded possibilities by incorporating diverse input modalities to facilitate multi-modal controllable generation. These recent unified architectures support multiple visual features as conditions. Emu2-Gen [49] uses an autoregressive model to predict the next tokens and uses a separated diffusion model to generate images. Instruct-Imagen [22] unifies image generation tasks together using multi-modal instructions. ACE [18] introduces the condition unit designed specifically for multi-modal tasks. OmniGen [55] uses an LLM as initialization and jointly models text and images within a single model to achieve unified representations across different modalities. UniReal [8] treats image-level tasks as discontinuous video generation, enabling a wide range of image generation and editing capabilities. In parallel developments, ChatDit [25] employs a multi-agent system for general-purpose, and interactive visual generation.

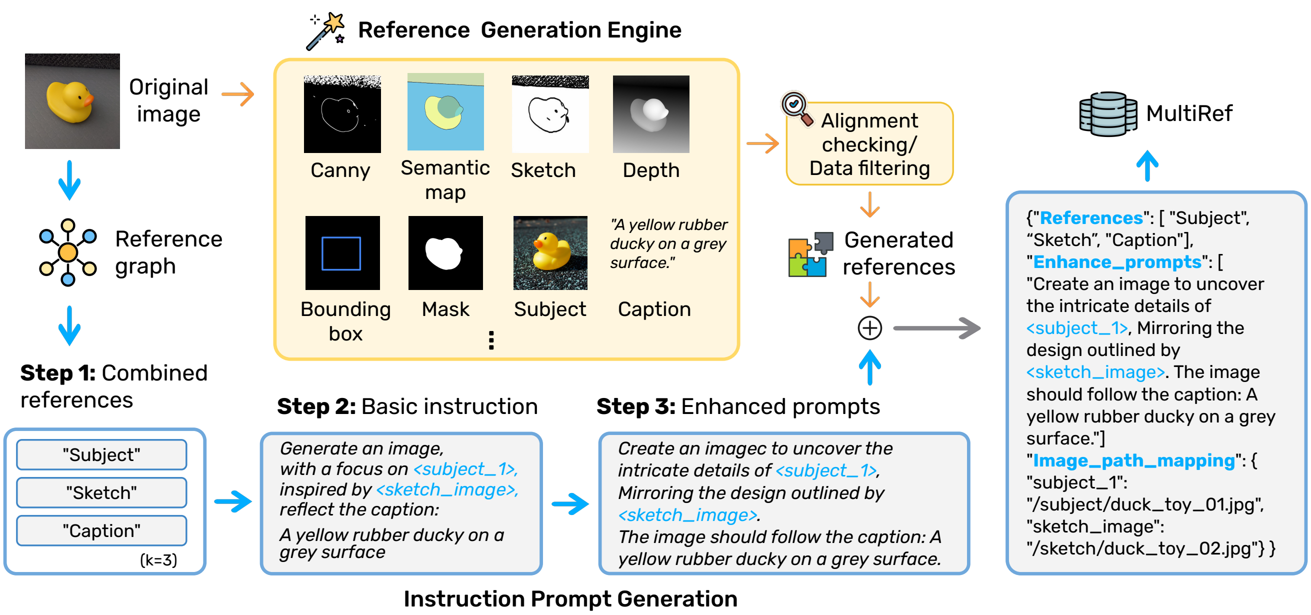

Figure 2. An overview of MULTIREF. It consists of reference generation (in yellow) and instruction prompt generation (in blue). First, various references (edges, semantics, depth) are extracted from an original image. Then, a basic instruction prompt is formed from selected compatible references. Finally, the enhanced prompt is integrated with references to construct a sample.

Dataset for Controllable Generation. Recent controllable image generation models have succeeded largely due to extensive training datasets like MultiGen-20M [41], which spans nine tasks across five categories with condition-specific instructions, while X2I dataset [55] incorporates flexible multi-modal instructions - yet these approaches still predominantly address single or dual conditions rather than complex, multi-reference combinations.

Previous work has established benchmarks for evaluating image generation, primarily focused on text-to-image quality and alignment [13, 16, 23, 24, 33] or image editing tasks [30, 48, 60]. Existing benchmarks like IDEA-Bench [32] and ACE benchmark [18] are limited in scope, with the former including images-to-image tasks but focusing primarily on editing operations like font transfer, while the latter only evaluates alignment with textual instructions—both failing to address complex scenarios involving multiple image references and their combinations.

3. MULTIREF-BENCH

To facilitate the evaluation and development of image generation models with multiple reference images, we introduce MULTIREF-BENCH, the first benchmark of its kind. Our approach combines real-world examples and synthetic data through a dual-pipeline methodology. The first pipeline gathers real-world tasks from publicly available internet sources, capturing authentic user needs and practical challenges. The second pipeline leverages traditional computer vision techniques to generate a broad and diverse set of conditional features. By integrating these two methodologies within a single dataset, we achieve a benchmark that is not only rooted in real-world applications but also expansive, diverse, and capable of evaluating models under a wide range of possible conditions.

3.1. Benchmark Overview

MULTIREF-BENCH consists of 1,990 examples. The first 1,000 examples represent real-world tasks sampled from the Reddit community `r/PhotoshopRequest`. This subreddit was selected for its diverse range of editing tasks, popularity, and active user engagement. The remaining 990 examples are test set that splited from 100,728 samples programmatically generation using REFBLEND — our custom framework for generating synthetic reference images, containing a diverse set of guidance signals, including depth maps, bounding boxes, art styles, and more to produce a wide array of conditional image generation scenarios.

3.2. Real-World Queries Collection

To develop a robust benchmark for evaluating conditional image generation models, we incorporate real-world, user-supplied tasks into our dataset. Authentic user interactions allow us to test models under diverse, practical conditions, capturing genuine challenges in real-world image editing. Following the methodology of RealEdit [50], we source real-world data from the `r/PhotoshopRequest` community on Reddit, a platform where users submit images and request professional-grade edits. These submissions

Table 1. Distribution of examples across different categories in real-world samples.

Category	Num.
Element Replacement	529
Element Addition	246
Spatial/Environment Modifications	111
Attribute Transfer	73
Style and Appearance Modifications	41
Total	1,000

cover a diverse array of editing tasks, authentically representing genuine user needs and practical challenges encountered in real-world image editing scenarios.

We collect 2,300 user queries, explicitly selecting tasks that require combining multiple input images to fulfill the requested edits. For each query, we gather all associated input images, the original text-based user instructions, and corresponding output images. To ensure data integrity and quality, each datapoint undergoes manual evaluation according to rigorous criteria. These criteria include verifying the necessity and appropriateness of each input image, assessing the logical coherence and relevance of instructions, and confirming accurate adherence to the instructions in the output image. In cases where multiple output images are provided for a single query, annotators select only one based on clarity, fidelity to the instruction, and overall quality.

To handle noisy human instructions and clearly specify references to individual images, we employ GPT-4o to generate structured prompts and detailed editing instructions. The model is explicitly guided to closely adhere to the original user requests while systematically incorporating image reference tokens (e.g., `<image1>`) to indicate elements of the edit corresponding to specific input images. All VLM-generated instructions subsequently undergo manual review to ensure clarity, consistency, and conformity to a standardized meta-prompt format. In instances where GPT-4o omits references to one or more input images, annotators manually correct and add the appropriate image tokens.

To provide insight into what edits are most commonly requested, we categorized each datapoint using the taxonomy structure proposed in OmniEdit [54]. The taxonomy comprises five categories: Element Replacement, Element Addition, Style and Appearance Modifications, Spatial/Environment Modifications, and Attribute Transfer. Each datapoint was processed using GPT-4o [39], following a standardized taxonomy prompt detailed in the Supplementary Material. The resulting distribution of edit types in our dataset is shown in Table 1.

After applying these rigorous quality standards and review processes, 45% of the collected data meet our criteria and are incorporated into the final benchmark dataset. This results in 1,000 examples, each comprising between two and six input images, a single structured instruction, and one output image as golden answer.

3.3. RefBlend: The Synthetic Data Engine

To construct an extensive benchmark, we develop a custom dataset generation engine, RefBlend, that employs a four-step process to automatically produce 100,728 diverse samples across 32 possible reference combinations. The process includes: (1) generating a comprehensive list of all potential reference conditioning (bounding boxes, depth maps, *etc.*), (2) programmatically produce a unique and exhaustive set of condition combinations based on compatible rules, (3) align multiple reference though a detailed text-based prompts, and (4) deploying a high-quality filtering pipeline to eliminate low-quality results. This structured approach ensures that only the most relevant and effective examples are included in the final dataset, resulting in a diverse and robust benchmark that covers a wide range of conditional image generation scenarios.

Step 1: Generate Reference Conditions. Given an original image, RefBlend leverages recent advanced models (*e.g.* Grounding Dino [43], Sam 2 [42] Depth Anything2 [58]), to synthesize a diverse set of conditioning inputs. These inclufr canny edges, semantic maps, sketches, depth maps, bounding boxes, masks, poses, art styles and subjects, along with textual captions generated by GPT-4o-mini [38]. These reference guidance types have proved themselves in controllable image generation in prior work [22, 41, 61, 62].

Our original images are sampled in a wide range from DreamBooth [45], CustomConcept101 [28], Subjects200K [51], WikiArt [46], Human-Art [27], StyleBooth [19] to X2I [55], which attach references about pose, subject, and art style within the dataset and for the diversity of metadata.

Step 2: Combining References. Not all references can be combined with each other. Some references are mutually exclusive, while others have specific dependencies that must be considered. To account for these complexities, we establish a set of visual reference compatibility rules. These rules define the valid combinations and dependencies among different image reference conditions. Following the rules ensures that only non-conflicting and meaningful reference combinations are used in dataset curation, avoiding redundancy. We establish three fundamental compatibility rules for image references:

(1) **Mutual Exclusivity of Global References:** References containing global information cannot be combined with each other, as this would result in information overlap. For example, Canny edge and sketch references, both capturing global structural information, are mutually exclusive because they provide a full structural view of the image.

(2) **Global-Local Information Incompatibility:** References with local information cannot be combined with those

Table 2. Reference Compatibility Matrix. Both rows and columns represent reference names. Yellow: Local Spatial Constraints. Green: Semantic Content Specification. Purple: Global Structural Guidance. Pink: Semantic Content Specification.

	Bounding box	Mask	Pose	Caption	Subject	Semantic map	Depth	Canny	Sketch	Art style
Bounding box	-	✗	✗	→	→	✗	✗	✗	✗	✓
Mask	✗	-	✗	→	→	✗	✗	✗	✗	✓
Pose	✗	✗	✓	✓	-	✗	✗	✗	✗	✓
Caption	✓	✓	✓	-	✓	✓	✓	✓	✓	✓
Subject	✓	✓	-	✓	✓	✓	✓	✓	✓	✓
Semantic map	✗	✗	✗	✓	✓	-	✗	✗	✗	✓
Depth	✗	✗	✗	✓	✓	✗	-	✗	✗	✓
Canny	✗	✗	✗	✓	✓	✗	✗	-	✗	✓
Sketch	✗	✗	✗	✓	✓	✗	✗	✗	-	✓
Art style	✓	✓	✓	✓	✓	✓	✓	✓	✓	-

✓: possible, i.e., the combination of row and column is feasible but does not depend on each other. ✗: impossible, i.e., the combination of row and column is invalid and cannot coexist. →: dependency, i.e., when the row is present, the corresponding column condition must also be met.

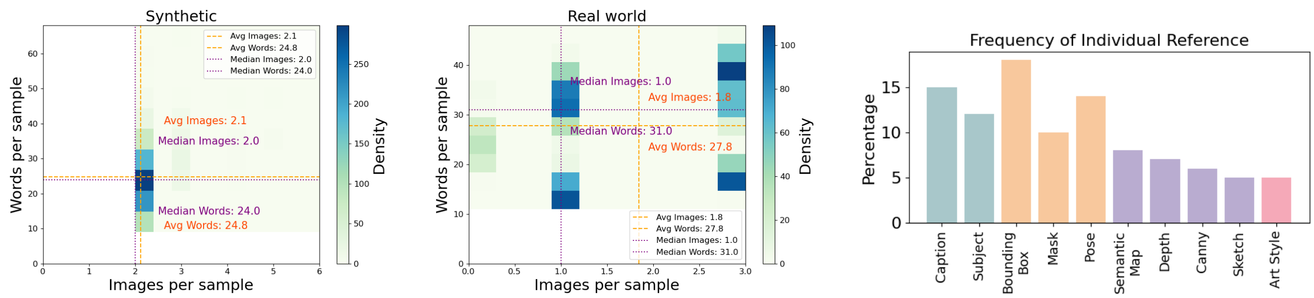

Figure 3. **Left, Middle:** Distribution analysis of textual content length and image count for synthetic and real-world parts. **Right:** Reference frequency in synthetic data.

containing global information to avoid redundancy. For instance, semantic maps (which provide a global understanding of image regions) cannot be combined with mask references (which localize specific objects) as this would create contradictory or redundant guidance.

(3) **Reference Dependencies:** Certain references have specific dependencies on others. For example, style transfer and caption references are universally compatible with all other references as they provide stylistic or descriptive context without overlapping spatial information. Conversely, spatial localization references (e.g., masks, bounding boxes) require semantic context (e.g., subject or caption) to accurately specify the desired content. A mask reference alone might indicate a region of interest, but without a semantic label or descriptive caption, the intended object or modification could remain ambiguous.

To ensure diversity and complexity within the dataset, we generate all possible combinations of 2, 3, and 4 references per instruction while strictly adhering to compatibility rules. These combinations evaluate models' capacity to integrate diverse guidance effectively.

Step 3: Generating Instructions. Using the valid reference combinations generated in Step 2, we create two types of prompts: structured and enhanced. Structured prompts are generated using a template-based approach that maps each reference type to a standardized phrase. For example, a depth reference might use the placeholder "*<depth_image>*" with associated phrases such as "*guided by the depth of <depth_image>.*" Caption references are appended with simple introductory phrases like "*following the caption:*". This method ensures that prompts are clear, consistent, and easy to parse, maintaining a straightforward format that models can readily interpret.

To broaden the scope and realism of our dataset, we transform structured prompts into more diverse and natural instructions using GPT-4o [39]. By applying different personas from Persona Hub [14], we vary the language, tone, and style of the prompts while maintaining the reference structure and intended content. This process not only enriches the prompts with creative and contextually relevant variations but also challenges models with a wide range of linguistic expressions and scenarios. The enhanced prompts, when combined with the generated references, result in a robust and versatile dataset suitable for comprehensive model evaluation.

Step 4: Filtering. Although the entire reference generation process is automated, advanced conditional generation models still produce errors in generated references, necessitating further inspection. After generating visual references, we apply a rule-based filter using metrics such as a confidence score threshold of 0.8 for the IoU (Intersection over Union) of semantic maps.

Table 3. Evaluation dimension and metrics of MULTIREF-BENCH for synthetic multi-ref generation. Rule: Golden standard for evaluation criteria. Model: We leverage a fine-tuned MLLM-as-a-judge for human-aligned semantic visual references evaluation.

Evaluation Dimension	Evaluation Aspect	Evaluation Criteria	Quantitative Metrics	Rule	Model
General Quality	Image Quality	Visual Fidelity	FID	✗	✔
	Visual Attractiveness	Aesthetic Appeal	CLIP Aesthetic Scores	✗	✔
Reference Fidelity	Bounding Box	Spatial Accuracy	IoU	✔	✗
	Semantic Map	Segmentation Accuracy	IoU	✔	✗
	Mask	Mask Alignmen t	IoU	✔	✗
	Depth Map	Depth Accuracy	MSE	✔	✗
	Canny Edge	Edge Preservation	MSE	✔	✗
	Sketch	Structural Fidelity	MSE	✔	✗
	Caption	Text-Image Alignment	CLIP Text-Image Score	✗	✔
	Pose ✻	Pose Accuracy	mAP	✔	✗
	Subject	Subject Consistency	CLIP Image Score	✗	✔
	Art Style	Style Consistency	CLIP Image Score	✗	✔
Instruction Following	Instruction Adherence	Instruction-Output Alignment	-	-	✔

✻ For pose, a single reference image may contain multiple instances (e.g., multiple poses merged in one reference image).

Table 4. Evaluating MLLM-as-a-Judge in scoring with cross-validated human-annotated ground truth. GPT-4o and 4o-mini aligns closely with human scores in overall assessment. Human-Human shows the alignment between human annotators.

Model	Image Quality				Instruction Following				Source Fidelity			
	Pearson	Spearman	MSE	MAE	Pearson	Spearman	MSE	MAE	Pearson	Spearman	MSE	MAE
Realistic												
Gemini-2.0-Flash	0.385	0.403	2.220	1.118	0.422	0.447	2.750	1.216	0.354	0.356	3.747	1.409
GPT-4o-mini	**0.466**	**0.466**	**1.676**	**0.986**	0.530	0.569	1.493	0.858	0.514	**0.518**	**1.193**	**0.733**
GPT-4o	0.432	0.420	2.486	1.223	**0.624**	**0.616**	**1.405**	**0.764**	**0.613**	0.513	1.216	0.736
Human-Human	0.589	0.573	1.611	0.936	0.665	0.590	1.152	0.720	0.571	0.441	1.473	0.824
Synthetic												
Gemini-2.0-Flash	0.369	0.347	2.078	1.052	0.627	0.592	1.662	0.855	0.588	0.574	2.057	0.960
GPT-4o-mini	**0.438**	**0.410**	**1.680**	**1.013**	0.632	0.552	**1.503**	0.870	0.616	0.615	2.173	1.140
GPT-4o	0.406	0.374	2.350	1.083	**0.668**	**0.608**	1.537	**0.843**	**0.659**	**0.626**	**1.573**	**0.860**
Human-Human	0.629	0.648	1.823	0.930	0.721	0.735	1.820	0.867	0.694	0.708	1.840	0.840

For more semantic-level visual references - such as subject, style, sketch, and canny - that do not provide confidence scores, we utilize a fine-tuned Qwen-2.5-2B-VL as an MLLM-as-a-Judge [5]. This model evaluates both the alignment between the original images and generated references and assesses their overall quality. Further details are provided in the Supplementary Material.

3.4. Evaluation

Our approach combines rule-based and model-based metrics to provide a comprehensive assessment of reference following capabilities across diverse conditions. The evaluation dimension and metrics of MULTIREF-BENCH are shown in Table 3. All evaluation metrics are finally normalized to a $[0, 1]$ range for consistency. For Reference Fidelity assessment, we calculate individual scores for each reference type, then derive the overall fidelity score by averaging across all references involved in a generation task.

Reference Fidelity. Reference Fidelity measures how accurately generated images preserve and incorporate the specific attributes, features, and characteristics from provided reference inputs. For the 10 reference types included in our benchmark, we employ specialized evaluation criteria and metrics tailored to each reference category. Spatial references (Bounding Box, Semantic Map, and Mask) are evaluated using IoU to quantify alignment accuracy. For structural references (Depth map, Canny edge, and Sketch), we calculate MSE to measure preservation fidelity. Pose accuracy is quantified with mAP. Semantic references receive specialized treatment: caption alignment is assessed using CLIP text-image scores [20], while subject consistency and style fidelity are evaluated using CLIP image-image scores. Notably, for aspects where rule-based quantitative metrics may not fully capture nuanced performance - particularly style consistency and subject fidelity - we supplement our evaluation with MLLM-as-a-Judge assessments by our fine-tuned model to provide complementary qualitative insights.

General Quality. General Quality assesses the overall visual quality and aesthetic appeal of generated images independent of reference fidelity. To evaluate this dimension

Table 5. Real-world image generation conditioned on multiple image references. Although today's image generative models produce high-quality outputs, most struggle with accurately following instructions and maintaining fidelity to source images. **IQ** - Image Quality, **IF** - Instruction Following, **SF** - Source Fidelity.

Model	Element Add.			Spatial Mani.			Element Rep.			Attribute Tran.			Style Modi.			Overall		
	IQ	IF	SF	IQ	IF	SF	IQ	IF	SF	IQ	IF	SF	IQ	IF	SF	IQ	IF	SF
Unified Model																		
Show-o	0.511	0.290	0.253	0.525	0.300	0.258	0.508	0.268	0.240	0.548	0.301	0.260	0.473	0.307	0.259	0.513	0.293	0.254
OmniGen	0.553	0.498	0.429	0.553	0.461	0.422	0.484	0.450	0.379	0.567	0.479	0.408	0.620	0.590	0.468	0.555	0.496	0.421
ACE	0.254	0.207	0.205	0.260	0.207	0.205	0.255	0.207	0.203	0.234	0.200	0.200	0.265	0.205	0.200	0.254	0.205	0.203
Compositional Framework																		
ChatDiT	0.629	0.390	0.345	0.643	0.411	0.352	0.643	0.434	0.360	0.682	0.466	0.395	0.688	0.522	0.424	0.657	0.445	0.375
Gemini+SD2.1	0.611	0.372	0.329	0.620	0.404	0.324	0.574	0.391	0.339	0.605	0.397	0.332	0.660	0.495	0.385	0.614	0.412	0.342
Claude+SD2.1	0.620	0.402	0.330	0.625	0.416	0.339	0.555	0.371	0.322	0.674	0.419	0.345	0.717	0.507	0.390	0.638	0.423	0.345
Gemini+SD3	0.764	0.590	0.478	0.729	0.589	0.453	0.725	0.540	0.452	0.715	0.556	0.452	0.785	0.640	0.485	0.744	0.583	0.464
Claude+SD3	0.744	0.578	0.454	0.751	0.586	0.456	0.675	0.497	0.408	0.745	0.556	0.441	**0.795**	0.629	0.478	0.742	0.569	0.447
Gemini+SD3.5	**0.786**	**0.615**	**0.500**	0.756	0.591	**0.473**	**0.759**	**0.558**	**0.459**	**0.789**	0.564	0.441	0.780	0.610	0.460	**0.774**	0.588	**0.467**
Claude+SD3.5	0.767	0.563	0.469	**0.777**	**0.598**	0.472	0.700	0.506	0.406	**0.789**	**0.625**	**0.466**	0.790	**0.654**	**0.498**	0.765	**0.589**	0.462
Ground Truth	0.711	0.797	0.712	0.751	0.780	0.748	0.651	0.714	0.624	0.772	0.722	0.692	0.780	0.820	0.756	0.733	0.767	0.706

Table 6. Comparison of model performance for multi-reference generation on the synthetic part. Although models perform well in overall assessment, they fail for generating image with multiple precise control signals.

Model	Overall Assessment			Image Quality		Reference Fidelity										
	IQ	IF	SF	FID↓	Aesthetic↑	AVG↑	BBox↑	Semantic Map↑	Mask↑	Depth Map↓	Canny Edge↓	Sketch↓	Caption↑	Pose*↑	Subject↑	Art Style↑
Unified Model																
Show-o	0.764	0.616	0.462	0.110	0.607	0.469	0.051	0.263	0.332	0.104	0.061	0.203	0.569	0.008	0.532	0.301
OmniGen	0.730	0.532	0.438	0.111	0.593	0.464	0.179	0.197	0.320	0.087	0.092	0.221	0.382	0.014	0.623	0.329
ACE	0.740	0.655	0.528	0.108	0.592	0.553	0.219	0.382	0.439	0.044	0.079	0.112	0.521	0.090	0.720	0.397
Compositional Framework																
ChatDiT	0.811	0.713	0.574	0.100	0.559	0.512	0.128	**0.176**	**0.393**	0.088	**0.065**	**0.207**	0.543	**0.018**	0.855	0.369
Claude + SD 2.1	0.812	0.726	0.572	**0.114**	0.612	0.488	**0.174**	0.132	0.292	0.203	0.080	0.230	0.547	0.005	0.817	0.424
Claude + SD 3	0.876	0.817	0.658	0.102	0.635	0.500	0.134	0.145	0.360	0.203	0.087	0.215	0.576	0.009	**0.859**	0.420
Claude + SD 3.5	**0.913**	**0.853**	**0.691**	0.111	**0.647**	**0.513**	0.124	0.147	0.358	0.082	0.082	0.213	0.573	0.009	0.858	**0.434**
Gemini + SD 2.1	0.791	0.708	0.547	0.113	0.615	0.477	0.161	0.133	0.255	0.202	0.092	0.239	0.550	0.003	0.791	0.406
Gemini + SD 3	0.856	0.804	0.639	0.103	0.635	0.507	0.141	0.135	0.368	0.083	0.121	0.216	**0.581**	0.008	0.840	0.414
Gemini + SD 3.5	0.893	0.839	0.676	0.111	0.646	0.510	0.132	0.130	0.371	**0.077**	0.096	0.216	0.579	0.008	0.845	0.422
Ground Truth	0.842	0.803	0.668	0.108	0.617	**0.709**	**0.410**	**0.772**	**0.893**	**0.000**	**0.000**	**0.000**	**0.584**	**0.149**	**0.869**	0.417

comprehensively, we employ two complementary metrics: FID [21] and CLIP aesthetic scores [47], to evaluate the image quality and creative aspects of the generated content.

Overall Assessment. For overall assessment, we follow Chen et al. [9] to leverage MLLM-as-a-Judge using GPT-4o-mini [38]. This approach evaluates overall Image Quality (IQ), Instruction Following (IF), and Source fidelity (SF) in a holistic manner. We validate the correlation of MLLM-as-a-Judge and human with a selected test set of 300 samples for either Realistic and Synthetic dataset. Our experiment in Table 4 reveals that GPT-4o-mini surpass other models in aligning with human. Therefore, we leverage GPT-4o-mini as our primary model for overall assessment.

4. Experiments and Analysis

4.1. Experiment Setups

Models. We conduct evaluations on three open-source unified image generation models: OmniGen [55], ACE [18], Show-o [56] [1]. For ACE and Show-o, we implement multi-turn dialogues to enable image generation with multiple references, incorporating one reference image per conversational turn. Additionally, we evaluate six compositional settings that specifically leverage Gemini-2.0-Flash [15] and Claude-3.7-Sonnet [2] as preceptors,[2] SD3 serves as the primary generator for dataset synthesis, with SD2.1 employed in ablation studies. Detailed configurations are available in the Supplementary Material.

4.2. Experiment Results

Compositional framework exceeds in image quality, while failing to maintain consistency on real-world cases. As shown in Table 5, SD3.5 combined with LLMs like Gemini and Claude achieves the highest scores among all tested approaches. Claude + SD3.5 attains exceptional image quality scores of 0.774 on average, occasionally surpassing ground truth. The clear progression in scores among three image generative models indicates that stronger image generative models achieve higher scores,

[1]Due to computation limitation, we do not employ Emu2-Gen [48].

[2]Given that GPT-4o participated in most of our experiments, we select alternative models for these compositional settings to avoid bias.

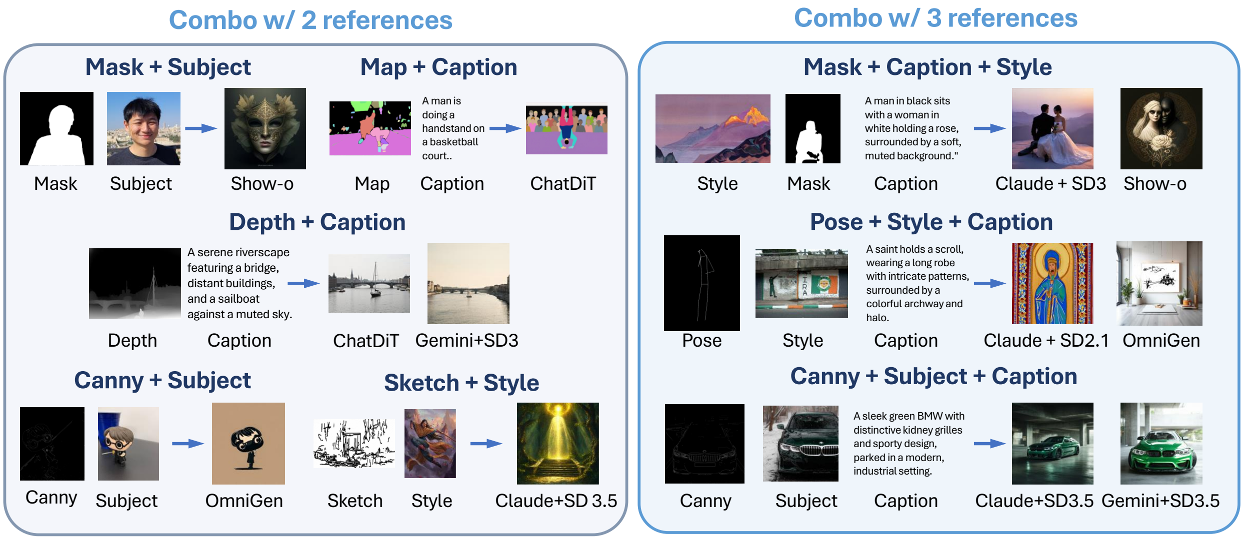

Figure 4. Case study of image generation conditioned on a combination of two and three references.

demonstrating that image quality significantly impacts evaluation metrics. However, all compositional frameworks consistently underperform in accurately following user's instruction and stay fidelity to source images in user's query. For instance, while ground truth has 0.767 and 0.706 for IF and SF respectively in the Overall category, Claude + SD3.5 only achieve 0.589 and 0.462, indicating that the separated preceptor-generator architecture fundamentally compromises the ability to faithfully interpret and execute complex visual instructions.

Unified models struggled with generation quality and handling real-world images. Although unified models theoretically end-to-end advantage contribute in maintaining consistency, they underperform in fidelity preservation as shown in Table 5. OmniGen's performance in various metrics even approaches some compositional frameworks that generate images with *state-of-the-art* diffusion models, demonstrating its effectiveness in balancing quality with instruction adherence. However, all models still fall short when comparing with golden answer (created with professional software), highlighting significant room for improvement in real-world image generation scenarios.

Controllable image generation from multiple references are challenging. As shown in Table 6, despite achieving high scores in Overall Assessment, substantial gaps remain in terms of strictly adhering to source fidelity—such as bounding box alignment, semantic map precision, and pose accuracy—when compared to the Ground Truth. Notably, the best-performing model, ACE, still exhibits considerable discrepancies in these aspects; however, it attains significantly better results in Bounding Box, Semantic Map, and Pose accuracy, with respective scores of 0.382, 0.439, and 0.720, underscoring the advantages of unified end-to-end methods for precise controllable image generation tasks. These observations suggest that unified architectures hold greater promise than traditional compositional frameworks employing separate modules, particularly in achieving fine-grained control over specific attributes, despite both approaches being able to produce visually appealing outputs.

Models fail when visual references are complexly mixed. We have discovered that when we input multiple visual references, even though these references do not conflict with each other, the model generates corrupted output and cannot produce correct images. Omnigen fails to output normal images when given black background bounding boxes or poses. Additionally, most of these unified models cannot generate images when the text prompt is removed. We believe that the robustness of current models regarding multi-image input for image generation still needs improvement.

5. Conclusion

Our work presents the first comprehensive investigation of image generation conditioned on multiple visual references, significantly expanding the boundaries of controllable image generation. Through developing a sophisticated synthetic data engine, we have constructed MULTIREF, a large-scale dataset for multi-reference image generation, from which we carefully curated a high-quality benchmark suite alongside a real-world application to MULTIREF-BENCH. Our systematic evaluation reveals that existing models, despite their claims of versatility, still face significant challenges when handling our multi-reference generation tasks. These findings provide valuable insights for the development of next-generation models that can more faithfully emulate the multi-reference creative processes inherent to human artistic expression, paving the way for more intuitive and expressive human-AI collaborative creation.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 674–681, 2024. 2

[2] Anthropic. Claude 3.5: A sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2024-09-04. 2, 7, 18

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 2

[4] Caroline Chan, Frédo Durand, and Phillip Isola. Learning to generate line drawings that convey geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7915–7925, 2022. 12

[5] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024. 2, 6, 13, 18

[6] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 2

[8] Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. *arXiv preprint arXiv:2412.07774*, 2024. 2

[9] Zhaorun Chen, Yichao Du, Zichen Wen, Yiyang Zhou, Chenhang Cui, Zhenzhen Weng, Haoqin Tu, Chaoqi Wang, Zhengwei Tong, Qinglan Huang, et al. Mj-bench: Is your multimodal reward model really a good judge for text-to-image generation? *arXiv preprint arXiv:2407.04842*, 2024. 7

[10] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 2

[11] 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. 1, 2, 18

[12] Flux. Black forest labs. https://blackforestlabs.ai/, 2024. 18

[13] Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Evaluating and improving text-to-vision generation with scene graph programming. *arXiv preprint arXiv:2412.08221*, 2024. 3

[14] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024. 5

[15] GeminiTeam. Gemini: A family of highly capable multimodal models, 2023. 2, 7, 18

[16] 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. 3

[17] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 1

[18] Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024. 2, 3, 7, 18

[19] Zhen Han, Chaojie Mao, Zeyinzi Jiang, Yulin Pan, and Jingfeng Zhang. Stylebooth: Image style editing with multimodal instruction. *arXiv preprint arXiv:2404.12154*, 2024. 1, 4, 13

[20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 2, 6

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

[22] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4754–4763, 2024. 2, 4, 11

[23] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 3

[24] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 3

[25] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Chen Liang, Tong Shen, Han Zhang, Huanzhang Dou, Yu Liu,

and Jingren Zhou. Chatdit: A training-free baseline for task-agnostic free-form chatting with diffusion transformers. *arXiv preprint arXiv:2412.12571*, 2024. 1, 2, 18

[26] Chengyou Jia, Changliang Xia, Zhuohang Dang, Weijia Wu, Hangwei Qian, and Minnan Luo. Chatgen: Automatic text-to-image generation from freestyle chatting. *arXiv preprint arXiv:2411.17176*, 2024. 1

[27] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 618–629, 2023. 4, 13

[28] 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. 4, 13

[29] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 2

[30] Tianle Li, Max Ku, Cong Wei, and Wenhu Chen. Dreamedit: Subject-driven image editing. *arXiv preprint arXiv:2306.12624*, 2023. 1, 3

[31] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[32] Chen Liang, Lianghua Huang, Jingwu Fang, Huanzhang Dou, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Junge Zhang, Xin Zhao, and Yu Liu. Idea-bench: How far are generative models from professional designing? *arXiv preprint arXiv:2412.11767*, 2024. 2, 3, 13

[33] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 3

[34] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 2

[35] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2

[36] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2

[37] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. In *European Conference on Computer Vision*, pages 93–110. Springer, 2024. 2

[38] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Accessed: 2024-09-04. 4, 7, 12

[39] OpenAI. Hello gpt-4o, 2024. Accessed: 2024-06-06. 4, 5, 18

[40] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 2

[41] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 11

[42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4, 12

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4, 11

[44] 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*, pages 10684–10695, 2022. 1, 2, 18

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 1, 4, 13

[46] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015. 4, 13

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 7

[48] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8871–8879, 2024. 3, 7, 18

[49] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 2

[50] Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. Realedit: Reddit edits as a large-scale empirical dataset for image transformations, 2025. 2, 3

[51] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024. 4, 13

[52] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 1, 2

[53] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. *Advances in Neural Information Processing Systems*, 37:128374–128395, 2024. 1

[54] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision, 2024. 4, 16

[55] 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. 2, 3, 4, 7, 13, 18

[56] 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. *arXiv preprint arXiv:2408.12528*, 2024. 2, 7, 18

[57] Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7754–7765, 2023. 2

[58] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 4, 12

[59] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 2

[60] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 4, 11

[62] 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, 2024. 1, 2, 4, 11

[63] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2

A. Details of Collecting MULTIREF-BENCH

We provide further details on the collection pipeline of MULTIREF-BENCH. Our dataset contains 100,728 samples in total, where 990 samples are split into test set for evaluation. See Table 7 for detailed statistics.

A.1. Reference Generation

We choose the guidance of image generation from prior work [22, 41, 61, 62], combining commonly used references, standardizing their names and adapting them to work with flexible input and output formats. Our final set of references includes edge maps (Canny), semantic maps, sketches, depth maps, bounding boxes, masks, poses, art styles and subjects, along with textual captions.

Bounding box. A bounding box is a small possible rectangular box that can completely enclose an object in an image, typically defined by the (x,y) coordinates of its top-left and bottom-right corners. We utilize phrase grounding in Grounded SAM2 [43] to identify and localize the main objects in a given image. The bounding box is visualized by drawing it on a black background of the same dimensions as the input image.

Mask. A mask is a binary image representation where the object of interest is separated from the background. It precisely outlines the shape and contour of the target object, creating a silhouette that exactly matches the object's boundaries rather than using a rectangular bounding box. We use Grounded SAM2 to generate masks, with one object corresponding to one mask. The mask is typically visualized as a binary image, where the background is represented by black pixels (value 0), and the object mask is represented by white pixels (value 1).

Pose. A pose refers to the spatial arrangement of key body parts (such as head, shoulders, elbows, wrists, hips, knees, and ankles) in a human figure, typically represented as a skeleton structure with joints and connections. The pose reference is visualized on a black background, with colored joints and connections highlighting the body's key positions and movements.

Semantic map. A semantic map, is a visual representation where each object class or semantic category is assigned a unique color or label, showing the location and boundaries of different semantic concepts in an image. We use AutomaticMaskGenerator in SAM2 [42] to generate the semantic map.

Depth map. A depth map is a grayscale image where each pixel's intensity represents the distance between the camera and the corresponding point in the scene. Typically, lighter/brighter pixels indicate points closer to the camera while darker pixels represent points that are farther away, creating a visual representation of the scene's 3D structure in a 2D format. We use Depth Anything V2 [58] to generate the depth map with default parameters.

Canny edge. A Canny edge map is a binary image that shows the boundaries and edges detected in an original image using the Canny edge detection algorithm. It identifies edges by looking for areas of rapid intensity change in the image, producing a clean, thin outline where white pixels represent detected edges against a black background. We use the Canny operation in OpenCV with thresholds in [100, 200].

Sketch. A sketch of an image is a simplified, line-based representation that captures the original image's essential contours and structural elements using only black lines on a white background. It focuses on preserving the key visual information while removing details like color, texture, and shading, similar to a hand-drawn outline. We use the line drawing method by Chan et al. [4] to generate the sketch reference, with contour_style and resize_and_crop process.

Art style. An art style of an image refers to the distinctive visual aesthetic, technique, or artistic treatment applied to transform the original image into a specific artistic rendering - such as watercolor, oil painting, cartoon and impressionist.

Subject. A subject reference image provides the main content or subject matter that needs to be transformed or recreated. It serves as the primary visual input that specifies what object or subject should be generated in the new image while maintaining its key characteristics and identity.

Caption. A caption of an image is a concise textual description that explains what is shown in the image, often describing the key subjects, actions, or notable elements present in the visual content. We use GPT-4o-mini [38] to describe the input image with prompts as follows.

Generate image caption

System prompt: You are a helpful assistant that can analyze images and provide detailed descriptions. Here is the image: [INSERT_IMAGES]

For subject-related images:
Describe this image in detail using no more than 20 words. Focus on the main subject in the image. Do not include any other unrelated information.

For other images:
Describe this image in detail using no more than 20 words. Do not include any other unrelated information.

Table 7. Distribution of Combinations by Count and Percentage generated by REFBLEND.

Combination	Count	(%)
caption+mask+subject	8,808	8.65
bbox+caption+subject	8,448	8.30
caption+depth+subject	8,304	8.15
caption+sketch+subject	6,456	6.34
caption+semantic map+subject	6,000	5.89
canny+caption+subject	5,424	5.33
caption+depth	4,032	3.96
caption+sketch+style	3,696	3.63
caption+semantic map	3,600	3.54
caption+sketch	3,528	3.46
canny+caption	3,456	3.39
canny+caption+style	3,384	3.32
caption+depth+style	3,384	3.32
caption+mask	3,048	2.99
caption+semantic map+style	2,856	2.80
bbox+caption	2,832	2.78
caption+pose+style	2,712	2.66
bbox+subject	2,400	2.36
bbox+caption+style	2,184	2.14
caption+pose	2,112	2.07
mask+subject	1,992	1.96
caption+mask+style	1,632	1.60
depth+subject	1,536	1.51
canny+subject	1,440	1.41
sketch+subject	1,416	1.39
caption+subject	1,248	1.23
semantic map+subject	1,104	1.08
canny+style	1,032	1.01
sketch+style	792	0.78
depth+style	792	0.78
mask+style	552	0.54
semantic map+style	528	0.52
Total	100,728	100.00%

A.2. Details of Metadata

Original images used for reference generation are from six datasets, as follows.

DreamBooth [45]. It is a collection of images used for fine-tuning text-to-image diffusion models for subject-driven generation. It includes 30 subjects from 15 different classes. Images of the subjects are usually captured in different conditions, environments, and under different angles. While DreamBooth offers subject references, it does not include art style or pose references.

Subjects200K [51]. It is a large-scale dataset containing 200,000 paired images. Each image pair maintains subject consistency while presenting variations in the scene context. The dataset does not include art style or pose references. We leverage subject references provided by the dataset itself.

CustomConcept101 [28]. It is a dataset consisting of 101 concepts with 3-15 images in each concept. The categories include toys, plushies, wearables, scenes, transport vehicles, furniture, home decor items, luggage, human faces, musical instruments, rare flowers, food items, pet animals. While it offers subject references, it does not include art style or pose references.

Human-Art [27]. It is a versatile human-centric dataset to bridge the gap between natural and artificial scenes. It includes twenty high-quality human scenes, including natural and artificial humans in both 2D representation and 3D representations. It includes 50,000 images in 20 scenarios, with annotations of human bounding box and human keypoints. From this dataset, we utilize two subsets: 2D_virtual_human and real_human, containing 22,000 and 10,000 images, respectively. Specifically, 2D_virtual_human provides art style and pose references while real_human provides pose references. Additionally, we leverage the art style and pose annotations provided within the dataset.

WikiArt [46]. WikiArt contains art paintings from 195 different artists. The dataset has 42,129 images for training and 10,628 images for testing. It does not include the subject reference or pose reference. We use images that share the same style as the art style references.

StyleBooth [19]. It is a high-quality style editing dataset accepting 67 prompt formats and 217 diverse content prompts, ending up with 67 different styles and 217 images per style. We use images that share the same style as the art style references.

X2I [55]. The entire dataset comprises approximately 0.1 billion images, including tasks of multi-modal instruction, subject-driven editing, in-context learning, computer vision and text-to-image generation. We use Web-Image, GRIT-Entity-New as metadata.

A.3. Data Filtering

To evaluate the complex outputs of free-form image generation, we assess both image quality and reference alignment using the MLLM-as-a-judge framework [5], which has gained widespread adoption in the field [32].

For each reference, the multimodal large language model examines both the original and generated images, evaluating alignment between them and assessing the quality of the generated reference (if applicable). The evaluation produces numerical scores on a 5-point scale (1-5), following specific scoring rubrics detailed below.

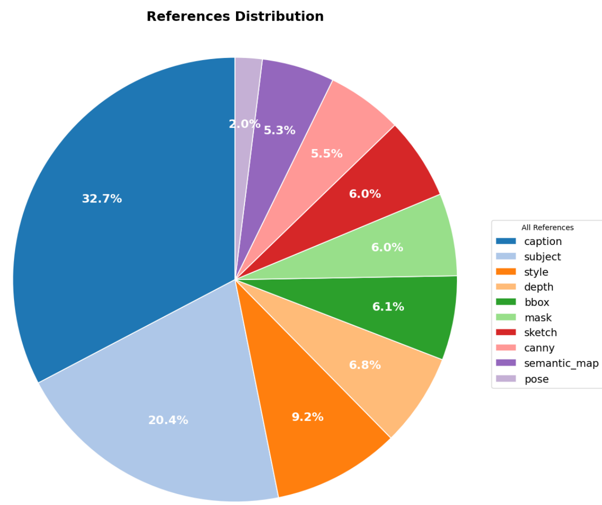

Figure 5. Reference Distribution

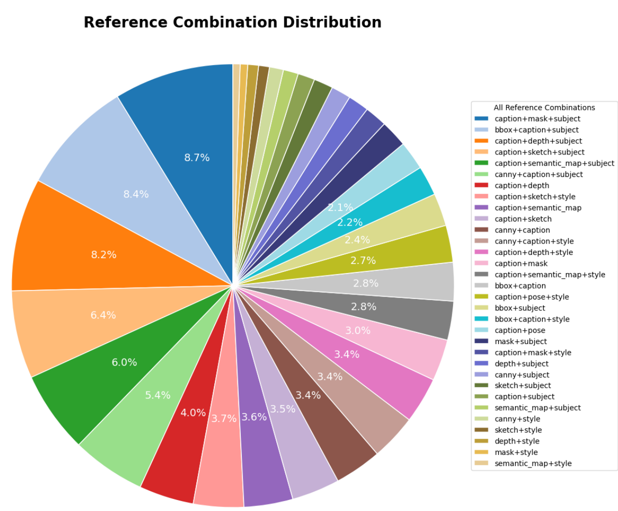

Figure 6. Reference Combination Distribution

Eval rubrics for canny

Definitions:
Canny Edge Map is a visual representation that highlights the edges and contours of objects in an image, where white lines represent detected edges and black represents non-edge regions.

Alignment:

- Not Aligned (Score 1) - Major object contours are unrecognizable or wrongly placed compared to the target image.
- Minimally Aligned (Score 2) - Few contours match the target image, with significant placement issues.
- Partially Aligned (Score 3) - Some major contours match while others are missing or misplaced.
- Mostly Aligned (Score 4) - Most main contours are recognizable and properly placed with minor misalignments.
- Well Aligned (Score 5) - Main object contours are recognizable and properly placed throughout the image.

Quality:

- Poor Quality (Score 1) - Excessive noise or breaks prevent object recognition entirely.
- Below Average Quality (Score 2) - Significant noise or breaks make most objects difficult to recognize.
- Average Quality (Score 3) - Key objects are recognizable despite moderate noise or breaks in contours.
- Good Quality (Score 4) - Main edges form clear object contours with minimal noise or breaks.
- High Quality (Score 5) - Main edges form recognizable object contours with the appropriate level of detail.

Eval rubrics for caption

Definitions:

Caption is a textual description that describes the content, context, objects, actions, or scene depicted in an image.

Alignment:

- Not Aligned (Score 1) - The caption describes elements that aren't present in the image, or fails to describe the main elements that are clearly visible.
- Minimally Aligned (Score 2) - The caption has minimal connection to the image content, with only one or two elements correctly identified.
- Partially Aligned (Score 3) -Some parts of the caption correctly describe the image while other described elements are missing or different, or the caption captures the general scene but misses key elements.
- Mostly Aligned (Score 4) - The caption describes most main elements and the overall scene with minor inaccuracies or omissions.
- Well Aligned (Score 5) - The caption accurately describes the main elements and scene in the image.

Eval rubrics for sketch

Definitions:

A sketch is a simplified, hand-drawn representation of an image, typically in black and white or grayscale, focusing on the main outlines and shapes of objects.

Alignment:

- Not Aligned (Score 1) - The basic object or scene structure is not captured at all.
- Minimally Aligned (Score 2) - Vague resemblance to the original image with major structural inaccuracies.
- Partially Aligned (Score 3) -The main concept is recognizable but with significant structural deviations.
- Mostly Aligned (Score 4) - Basic shapes and composition generally match with minor proportional variations.
- Well Aligned (Score 5) - The basic shapes and composition match accurately to the original image.

Quality:

- Poor Quality (Score 1) - Excessive noise or unclear lines make it difficult to interpret the intended subject.
- Below Average Quality (Score 2) - Substantial noise or rough elements that significantly detract from the subject.
- Average Quality (Score 3) -The sketch shows the subject but includes noticeable noise, scattered marks, or rough elements while maintaining recognizable forms.
- Good Quality (Score 4) - Clear lines with minimal noise that effectively represent the subject.
- High Quality (Score 5) - Clean, clear lines that effectively convey the subject with minimal noise or distraction.

Eval rubrics for semantic map

Definitions:

A semantic map is a visual representation where an image is divided into distinct regions to represent different objects, areas, or elements of the scene, using any colors or styles to distinguish between regions.

Alignment:

- Not Aligned (Score 1) - The basic scene structure or main objects are unrecognizable.
- Minimally Aligned (Score 2) - Only a few elements are recognizable, with significant missing or misplaced components.
- Partially Aligned (Score 3) - Some key elements are recognizable but others are missing or unclear.
- Mostly Aligned (Score 4) - Most elements capture recognizable objects and scene layout with minor inaccuracies.
- Well Aligned (Score 5) -The map captures recognizable objects and scene layout appropriately (simplified shapes are acceptable, textures and fine details not required).

Quality:

- Poor Quality (Score 1) - Semantic regions are too sparse or scattered to identify main objects; regions are too minimal to understand scene content.
- Below Average Quality (Score 2) - Main elements are barely distinguishable with significant noise, artifacts, or fragmented segments that impair understanding.
- Average Quality (Score 3) - Main elements are clearly visible but with noticeable noise/artifacts or scattered segments, while still maintaining recognizable object shapes.
- Good Quality (Score 4) - Key objects/regions are well-defined with limited noise or artifacts; segmentation is generally clean with only minor issues.
- High Quality (Score 5) - Main objects/regions are clearly visible and distinguishable, with clean segmentation of major elements; minimal artifacts or noise around edges.

Eval rubrics for mask

Definitions:

Mask Image is a binary image where white regions indicate areas of interest or target regions for object placement/generation, while black regions represent background or non-target areas.

Alignment:

- Not Aligned (Score 1) - Main parts of the main object are not covered by the mask, or the mask position doesn't correspond to the object location.
- Minimally Aligned (Score 2) - The mask covers only a small portion of the main object or is significantly misplaced.
- Partially Aligned (Score 3) - The mask covers most but not all of the main object, or if positioning is noticeably off.
- Mostly Aligned (Score 4) - The mask covers the main object with minor positioning issues or slight shape inaccuracies.
- Well Aligned (Score 5) - The mask captures the general outline and position of the main object accurately.

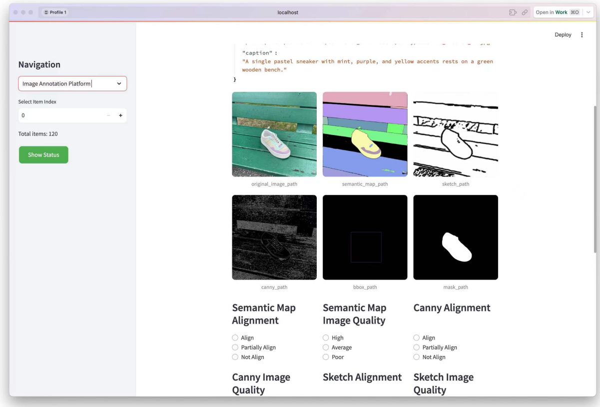

Figure 7. Human Annotation Platform

Reference with scoring under 3 will be filtered in the checking process. Pose, subject, and art style references are manually verified, as they are provided by the dataset and contain minimal annotation errors.

A.4. Human Annotation

The annotation process was conducted by three independent evaluators: two authors of this paper and one volunteer. Recognizing that annotator diversity is essential for minimizing bias and maximizing dataset reliability, we selected annotators with varying demographic characteristics (gender, age, and educational background) while ensuring all possessed domain expertise in image generation evaluation.

To establish annotation consistency and objectivity, all evaluators underwent comprehensive training sessions before beginning the task. These sessions included detailed tutorials on objective image assessment techniques, familiarization with reference rubrics, and instruction on the specific criteria used in our Score Evaluation framework. This preparatory process ensured methodologically sound and comparable annotations across all dataset entries.

The annotation platform is shown in Figure 7.

CVPR
#XXX

B. Benchmark Construction

B.1. Real-world

Taxonomy creation. We adopted the taxonomy structure introduced in OmniEdit [54] to categorize the types of edits represented in our benchmark. We utilized GPT-4o with the following prompt to generate the taxonomy for our dataset.

Prompt of generating taxonomy for real-world queries

You are tasked with classifying image editing instructions into one of the following 5 categories:
1. Element Replacement - Face swaps - Object substitutions - Background replacements - Text replacements - Component swaps (wheels, screens, etc.)
2. Element Addition - Adding people to scenes - Adding objects to environments - Adding details or elements to objects - Adding text or graphics - Adding visual effects
3. Style and Appearance Modifications - Color adjustments - Lighting modifications - Artistic style transfers - Texture changes - Visual quality enhancements
4. Spatial/Environment Manipulations - Repositioning elements - Combining multiple images into layouts - Changing scale or proportion - Adjusting orientation or alignment - Creating composite images
5. Attribute Transfers - Transferring expressions between faces - Applying visual characteristics across images - Maintaining specific features while changing others - Matching visual properties (lighting, color) - Transferring specific details while preserving context

---

Given the following image editing instruction, classify it into exactly one of these 5 categories. Respond with a JSON object with a single key "category" and the value being the category number (1-5).

To produce the meta-style prompts from noisy user instructions, we used the prompt with gpt-4o. We supplied all input images, corresponding output image as well as original user instructions.

GPT-4o prompt for rewriting instructions

You are an expert at image editing. Your job is to write a prompt that would help machine learning models to edit images.

I'm showing you:
1. First, the INPUT IMAGE(S) that the user wants to edit.
2. Then, the user's ORIGINAL INSTRUCTION (which might be noisy or unclear).
3. Finally, the OUTPUT IMAGE after editing.

Based on comparing these, please:
1. Infer what specific edit was performed
2. Write a clear, precise prompt that would help an AI model achieve this exact edit

Your prompt should follow this format:
"Edit image `<image1>` by [specific editing instruction using clear terminology]"

Here are some examples of good output prompts:
- "Edit image `<image1>` by taking the person from `<image2>`, person from `<image3>` and adding them to `<image1>`."
- "Edit image `<image2>` by transferring the background from `<image1>` and replacing the person with the person from `<image3>`"
- "Edit image `<image1>` by faceswapping the person from `<image2>` into `<image1>`"

Now, analyze the following:
ORIGINAL INSTRUCTION: {{description}}

Please provide a well-structured, clear editing prompt that precisely describes the transformation shown in the images.

CVPR
#XXX

Example 1:
Instruction: Edit image **<image3>** by replacing the face of the left caroler with the face from **<image1>** and the face of the right caroler with the face from **<image2>**.

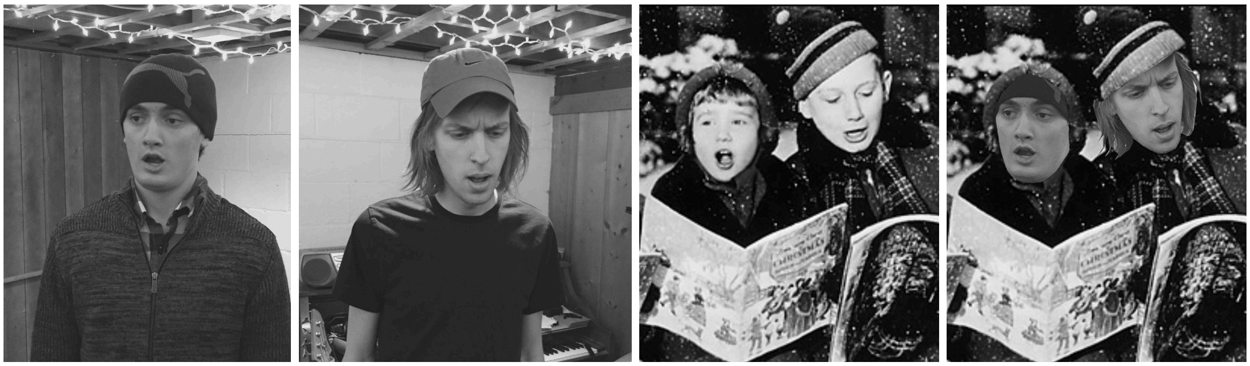

Image1 Image 2 Image3 Sourced Ground Truth

Example 2:
Instruction: Edit image **<image1>** by adding the window pattern from **<image2>** into the sky area on the right side of **<image1>**, blending it seamlessly with the existing snow and lighting effects.

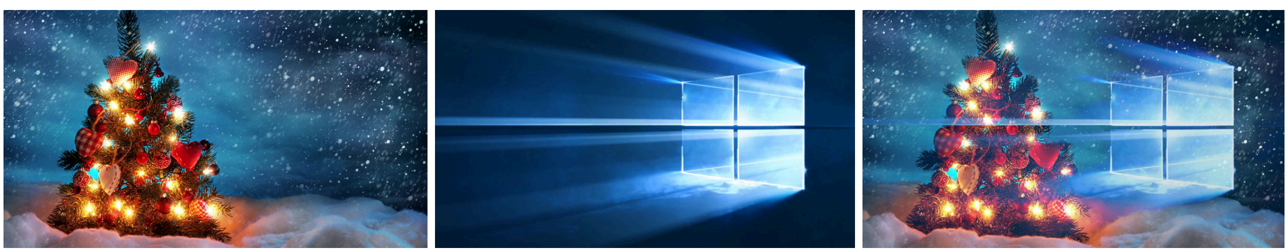

Image1 Image 2 Sourced Ground Truth

Example 3:
Instruction: Edit image **<image1>** by placing the child from **<image2>** into the arms of the person, ensuring the child appears to be sitting naturally and is proportionate to the person holding them.

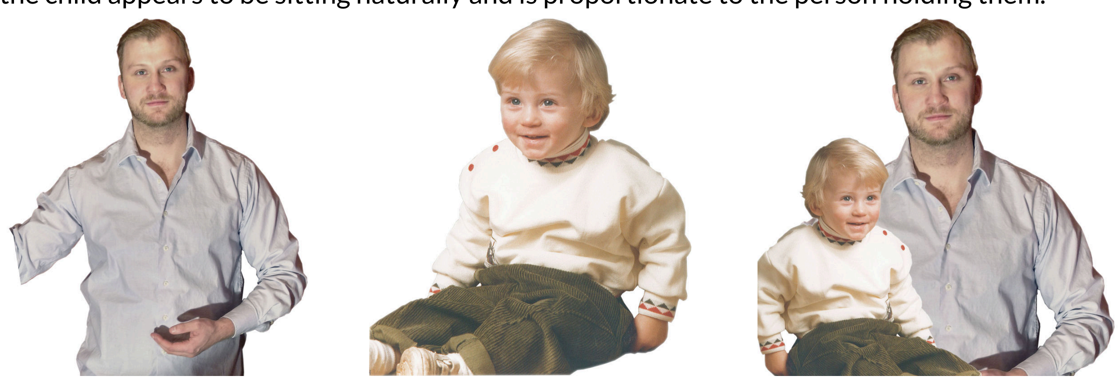

Image1 Image 2 Sourced Ground Truth

Figure 8. We visualize example datapoints in the realworld half of our benchmark. These examples are sourced from Reddit's r/PhotoshopRequest community.

B.2. Details of Fine-tuning MLLM-as-a-Judge for Filtering

For more semantic-level visual references such as subject, style, sketch, canny, and edge that do not provide confidence scores, we establish a fine-tuned MLLM-as-a-Judge [5] that verifies both the alignment between original images and generated references and their quality. Specifically, we collect a subset with 6,400 original images and their corresponding references, constructing them into (image, reference) pairs and subsequently collect cross-validated human scoring from 1-5 for alignment and quality score individually. Finally, we split it into train/test sets, each with 16,590 and 1,750 samples, and fine-tune Qwen-2.5-2B-VL. Our evaluation of the fine-tuned model with human-annotated scores as ground truth across Pearson similarity and MAE in Table 8 reveals close alignment with human annotators, validating it as a good judge for filtering.

C. Details of Experiments Setups

C.1. Model Settings

In this section, we will introduce the hyper-parameters of image generative models to facilitate experiment reproducibility and transparency. All our experiments were conducted on a server equipped with two A800 and two 4090 GPUs.

Open-source Unified Models. We employed four open-source unified models. All hyper-parameters are detailed as follows:

- **OmniGen [55].** We set height=1024, width=1024, guidance_scale=2.5, img_guidance_scale=1.6, seed=0 as default settings.
- **ChatDit [25].** We use the images-to-image API call provided in the GitHub.
- **ACE [18].** We use the ACE-0.6B-512px as ACE-chat model for multi-reference image generation in multi-turn. We set sampler='ddim', sample_steps=20, guidance_scale=4.5, guide_rescale=0.5.
- **Show-o [56].** We use multi-turn dialogue for multi-reference image generation. We set guidance_scale=1.75, generation timesteps=18, temperature=0.7, resolution: 256×256.

As reported in GitHub, Emu2-Gen [48] needs at least 75GB of memory. Due to the limitation of computation, it is not employed in our experiments.

Other Models. We utilize three proprietary models, GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-pro-latest as multimodal preceptors and Flux-dev, SD3, SD2.1 as image generators, with detailed settings as follows:

- **Gemini-1.5-pro-latest [15].** Temperature=1, top_p=0.95.
- **Claude-3.5-Sonnet [2].** Temperature=0.9.
- **GPT-4o [39].** Temperature=1, top_p=1.
- **Flux1-dev [12].** guidance scale=3.5, num inference steps=50.
- **Stable Diffusion 3 [11].** guidance scale=7.0, num inference steps=28.
- **Stable Diffusion 2.1 [44].** guidance scale=7.5, num inference steps=25.

C.2. Prompt Template

For each reference, we generate 10 structured basic instructions, as shown below.

Basic instructions for Art Style

- Inspired by the essence of ⟨style_image⟩, this reflects its distinctive flair
- Crafted in the characteristic tone of ⟨style_image⟩
- Modeled with the unique influence of ⟨style_image⟩
- Echoing the artistic spirit of ⟨style_image⟩
- Infused with the signature style of ⟨style_image⟩
- Reflecting the aesthetic nuances of ⟨style_image⟩
- A reinterpretation influenced by ⟨style_image⟩
- Harmonizing with the thematic essence of ⟨style_image⟩
- Inspired by and shaped in the vein of ⟨style_image⟩
- Capturing the creative vision embodied by ⟨style_image⟩

Basic instructions for Sketch

- Following the sketch of ⟨sketch_image⟩, this mirrors its essence.
- Designed in alignment with the sketch of ⟨sketch_image⟩.
- Echoing the framework drawn by ⟨sketch_image⟩.
- Guided by the outline of ⟨sketch_image⟩, it retains its authenticity.
- Reflecting the initial strokes of ⟨sketch_image⟩.
- Infused with the skeletal form of ⟨sketch_image⟩.
- Shaped under the influence of ⟨sketch_image⟩'s sketch.
- Structured around the design of ⟨sketch_image⟩.
- Capturing the structural integrity of ⟨sketch_image⟩.
- Crafted to reflect the framework of ⟨sketch_image⟩.

Table 8. Our fine-tuned MLLM-as-a-Judge scoring align closely with human preferences in assessing semantic-level visual reference.

Condition Type	Subject		Depth		Caption		Mask		Style		Sketch		Semantic Map	
	Pearson	MAE	Pearson	MAE	Pearson	MAE	Pearson	MAE	Pearson	MAE	Pearson	MAE	Pearson	MAE
Sample Size	158		358		169		74		158		276		154	
GPT-4o-mini	0.759	0.779	0.367	1.592	0.782	0.580	0.390	1.662	0.457	0.949	0.490	1.290	0.404	1.188
Qwen-2.5-2b-vl-zs	0.231	1.475	0.044	1.631	0.864	0.491	0.051	1.987	0.173	2.671	0.239	1.446	0.209	1.338
Qwen-2.5-7b-vl-zs	0.694	0.892	0.148	1.757	0.869	0.515	0.495	1.757	0.293	1.171	0.270	1.515	0.618	1.117
Qwen-2.5-2b-vl-ft (ours)	**0.726**	**0.722**	**0.581**	0.944	0.853	0.509	0.386	1.216	0.402	0.949	0.567	1.007	0.622	1.124

Basic instructions for Depth

- Following the depth of ⟨depth_image⟩, this delves into its essence.
- Inspired by the dimensionality of ⟨depth_image⟩, it captures its core.
- Reflecting the profound layers of ⟨depth_image⟩.
- Echoing the spatial depth of ⟨depth_image⟩, it retains its integrity.
- Infused with the visual perspective of ⟨depth_image⟩.
- Guided by the textured depth of ⟨depth_image⟩.
- Structured to align with the depths captured by ⟨depth_image⟩.
- Modeled after the layered depth of ⟨depth_image⟩.
- Harmonizing with the multi-dimensional feel of ⟨depth_image⟩.
- Crafted to embrace the depth portrayed by ⟨depth_image⟩.

Basic instructions for Canny

- Following the edge of ⟨canny_image⟩, this captures its sharpness.
- Inspired by the contours of ⟨canny_image⟩, it traces its form.
- Reflecting the defined edges of ⟨canny_image⟩.
- Echoing the precision lines of ⟨canny_image⟩, it retains its clarity.
- Infused with the sharp boundaries of ⟨canny_image⟩.
- Guided by the linear features of ⟨canny_image⟩.
- Structured to follow the contours highlighted by ⟨canny_image⟩.
- Modeled after the crisp edges of ⟨canny_image⟩.
- Harmonizing with the boundary lines of ⟨canny_image⟩.
- Crafted to reflect the edge details of ⟨canny_image⟩.

Basic instructions for Semantic Map

- Following the semantic map in ⟨semantic_image⟩, this aligns with its meaning.
- Inspired by the structure of ⟨semantic_image⟩, it conveys its intent.
- Reflecting the mapped semantics of ⟨semantic_image⟩.
- Echoing the visual language of ⟨semantic_image⟩, it captures its essence.
- Infused with the meaningful contours of ⟨semantic_image⟩.
- Guided by the symbolic layout of ⟨semantic_image⟩.
- Structured around the semantics depicted in ⟨semantic_image⟩.
- Modeled to align with the conceptual map of ⟨semantic_image⟩.
- Harmonizing with the thematic essence of ⟨semantic_image⟩.
- Crafted to reflect the semantic details of ⟨semantic_image⟩.

Basic instructions for Bounding Box

- Following the bounding box in ⟨bbox_image⟩, this outlines its structure.
- Inspired by the box constraints of ⟨bbox_image⟩, it defines its scope.
- Reflecting the encapsulated regions of ⟨bbox_image⟩.
- Echoing the boundary lines of ⟨bbox_image⟩, it retains its precision.
- Infused with the spatial framework of ⟨bbox_image⟩.
- Guided by the rectangular limits of ⟨bbox_image⟩.
- Structured to follow the defined areas in ⟨bbox_image⟩.
- Modeled after the bounding parameters of ⟨bbox_image⟩.
- Harmonizing with the enclosed regions of ⟨bbox_image⟩.
- Crafted to reflect the boundary specifications of

⟨bbox_image⟩.

Basic instructions for Single Mask

- Following the mask in ⟨mask_image⟩, this captures its shape.
- Inspired by the masked outline of ⟨mask_image⟩, it defines its form.
- Reflecting the contours covered by ⟨mask_image⟩.
- Echoing the masked regions of ⟨mask_image⟩, it retains its detail.
- Infused with the coverage specified by ⟨mask_image⟩.
- Guided by the spatial coverage of ⟨mask_image⟩.
- Structured to align with the masked features in ⟨mask_image⟩.
- Modeled after the outlined mask of ⟨mask_image⟩.
- Harmonizing with the masked boundaries of ⟨mask_image⟩.
- Crafted to reflect the regions defined by the mask in ⟨mask_image⟩.

Basic instructions for Pose

- Following the pose in ⟨pose_1⟩, this mirrors its stance.
- Inspired by the posture captured in ⟨pose_1⟩, it reflects its form.
- Reflecting the alignment depicted in ⟨pose_1⟩.
- Echoing the position shown in ⟨pose_1⟩, it retains its essence.
- Infused with the dynamic structure of ⟨pose_1⟩.
- Guided by the articulated motion of ⟨pose_1⟩.
- Structured around the pose outlined in ⟨pose_1⟩.
- Modeled to replicate the position in ⟨pose_1⟩.
- Harmonizing with the posture embodied in ⟨pose_1⟩.
- Crafted to reflect the expressive pose of ⟨pose_1⟩.

Basic instructions for Subject

- featuring ⟨subject_1⟩.
- showcasing ⟨subject_1⟩.
- focusing on ⟨subject_1⟩.
- while emphasizing ⟨subject_1⟩
- with a focus on ⟨subject_1⟩.
- centered on ⟨subject_1⟩.
- highlighting ⟨subject_1⟩.
- to better display ⟨subject_1⟩.
- while emphasizing ⟨subject_1⟩.
- to reveal finer details of ⟨subject_1⟩.

We use the prompt Diversity enhancement to write enhanced instructions, shown as below.

Diversity enhancement

You will adopt the persona of selected_persona. You will be given a text and your task is to rewrite and polish it in a more diverse and creative manner that reflects the persona's style. Do not include any direct references to the persona itself.
You may alter sentence structure, wording, and tone.
Do not modify text enclosed in angle brackets ”.
If there is a 'caption:' section in the text, do not change anything following 'caption:'
Here is the text: basic_instruction
Please provide the revised text directly without any additional commentary.