OpenSubject: Leveraging Video-Derived Identity and Diversity Priors for Subject-driven Image Generation and Manipulation

Yexin Liu[1,2] Manyuan Zhang[2,*] Yueze Wang[3] Hongyu Li[2]
Dian Zheng[2] Weiming Zhang[4] Changsheng Lu[1] Harry Yang[1†]

[1]HKUST, Hong Kong, China [2]Meituan, China
[3]Independent Researcher, China [4]HKUST(GZ), Guangzhou, China
[*]Project Leader [†]Corresponding Author
yliu292@connect.ust.hk, harryyang.hk@gmail.com
GitHub Link: https://github.com/LAW1223/OpenSubject

Abstract. Subject-driven image generation and manipulation are fundamental for personalized content creation, such as identity-preserving portrait synthesis, multi-character storytelling, and controllable photo editing. However, existing models still struggle with multi-reference settings, often showing identity drift and unstable editing quality. A key bottleneck is data: current resources are often limited in scale, diversity, and consistency of subject-level details, especially for unified support of both generation and manipulation. To address this gap, we introduce **OpenSubject**, a video-derived large-scale corpus for subject-driven generation and manipulation. Our pipeline leverages cross-frame identity priors through four key stages: (i) **Video Curation**. We apply resolution and aesthetic filtering to obtain high-quality clips. (ii) **Cross-Frame Subject Mining and Pairing**. We utilize vision language model (VLM)-based category consensus, local grounding, and diversity-aware pairing to select image pairs. (iii) **Identity-Preserving Reference Image Synthesis**. We introduce segmentation map-guided outpainting to synthesize input images for subject-driven generation and box-guided inpainting to generate input images for subject-driven manipulation, together with geometry-aware augmentations and irregular boundary erosion. (iv) **Verification and Captioning**. We utilize a VLM to validate synthesized samples, re-synthesize failed samples based on stage (iii), and then construct short and long captions. Furthermore, we introduce a benchmark covering both subject-driven generation and subject-driven manipulation, evaluated with a VLM-based judge on identity fidelity, prompt adherence, and consistency. Extensive experiments demonstrate that training on OpenSubject significantly enhances performance in complex scenes.

Keywords: Subject-driven · Image generation · Image manipulation

1 Introduction

Subject-driven image generation aims to synthesize realistic images of specified subjects conditioned on text prompts while preserving identity [7, 31]. This

Fig. 1: *OpenSubject* examples for subject-driven image generation and manipulation in both single- and multi-subject settings. The dataset covers humans, objects, and cartoon characters across indoor/outdoor scenes and diverse viewpoints, emphasizing identity fidelity and contextual diversity.

capability underpins applications in personalized content creation [22], portrait or character re-rendering [12], and interactive editing [10, 36]. With the advent of Diffusion Transformers (DiTs) [28], methods in this area have achieved substantial progress, particularly in the single-subject setting.

Despite these advances, two practical failures remain common. First, in subject-driven generation, fine-grained identity details (e.g., face structure and local texture) are often inconsistent across outputs. Second, in subject-driven manipulation (e.g., identity replacement), models frequently fail to preserve target-subject consistency and also modify non-target pixels that should remain unchanged. We argue that these failures are largely data-driven. Effective training requires paired supervision that is (i) identity-consistent and (ii) diverse in viewpoint and context. Existing subject-driven data pipelines (Tab. 1) still have clear limitations. *Synthesis-based* pipelines [33, 40, 46] are scalable, but often suffer from weak fine-detail consistency due to generator-induced artifacts and identity drift. *Retrieval-based* pipelines [44] better preserve realism, but are difficult to scale and provide limited control in open-ended multi-subject scenes. More importantly, most prior pipelines are designed for generation-only supervision and do not explicitly support building large-scale subject-driven manipulation datasets.

Our key idea: use video as scalable, subject-consistent supervision. We revisit video as a powerful yet underexploited supervision source for subject-driven modeling. Unlike web retrieval, videos provide repeated observations of the *same* subject under naturally varying viewpoints, lighting, motion, and interactions; unlike pure synthetic pipelines, this identity signal is grounded in real

Table 1: Comparison of subject-driven image generation and manipulation datasets. Counts denote the number of *paired input samples*.

Dataset	Paired Single-subject Input Samples	Paired Multi-subject Input Samples	Subject-driven Manipulation	IP Source	General Objects	Diverse Contexts
Echo-4o-Image [46]	✗	73k	✗	Synthesis	✓	✗
UNO-1M [40]	1M	✗	✗	Synthesis	✓	✗
Subjects200K [33]	200k	✗	✗	Synthesis	✓	✗
MultiID-2M [44]	✗	500k	✗	Retrieval	✗	✗
MICo-150k [37]	✗	150k	✗	Synthesis	✓	✓
OpenSubject (ours)	**748k**	**1752k**	✓	**Video**	✓	✓

visual evidence rather than a generator prior. This motivates a different recipe: anchor *who/what* the subject is in real video, and use controlled synthesis only to construct task-specific *inputs*. Concretely, we do not synthesize target outputs; instead, we synthesize diverse, challenging inputs (e.g., occluded or incomplete observations) from real frames and train models to recover subject-consistent results. This design reduces identity drift, improves controllability, and scales to open-ended, multi-subject scenarios (Fig. 1).

Building on this insight, we introduce **OpenSubject**, a large-scale corpus for *open-ended* subject-driven image generation and manipulation, built with a four-stage pipeline that addresses the above data bottlenecks. In particular, we focus on two practical challenges: selecting cross-frame pairs that maximize diversity while preserving subject consistency, and increasing context variation beyond near-identical within-clip backgrounds.

To address these challenges, we introduce **OpenSubject**, a dataset constructed via a four-stage pipeline with strict scene-level splits. (i) *Video curation*: we collect large-scale open-source videos and apply resolution and aesthetic filters to obtain high-quality, subject-persistent clips. (ii) *Cross-frame subject mining and pairing*: we sample candidate frames, enforce clip-level subject consensus with a vision–language model (VLM), perform instance-level local verification with Grounding-DINO [20] and the VLM, and select a maximally diverse frame pair using DINOv2 embeddings [27]. (iii) *Identity-preserving reference image synthesis*: we use mask-guided outpainting to construct cross-frame inputs for subject-driven generation, complemented by geometry-aware augmentations and irregular boundary erosion; we also introduce an editing task, *subject-driven manipulation*, defined as replacing the target subject in the input image with the referenced identity while preserving all non-target content, for which we synthesize inputs via box-guided inpainting. (iv) *Verification and captioning*: we validate synthesized samples with a VLM, re-synthesize failures based on stage (iii), and generate both short and long captions. The resulting corpus (2.5M samples, 4.35M images) supports both subject-driven generation and manipulation (see Fig. 1). OpenSubject is a video-grounded yet synthetic-paired corpus: real videos define the subject identities, poses, and multi-view context, while FLUX.1 Fill [dev] [16] in-/out-painting is used to synthesize the training inputs.

We further establish an open-ended benchmark that spans subject-driven generation and manipulation, comprising four sub-tasks and evaluated with a rubricized VLM judge (GPT-4.1 [24]), reporting identity fidelity, prompt adherence, manipulation fidelity, and background consistency. Extensive experiments

demonstrate that models trained on OpenSubject achieve stronger identity preservation and manipulation fidelity, particularly in multi-subject settings.

Our main contributions are as follows: (i) **Large-scale, video-derived corpus for subject-driven generation and manipulation**. We present OpenSubject, a 2.5M-sample (4.35M-image) dataset for subject-driven image generation and manipulation, covering both single- and multi-subject settings. (ii) **Scalable video-based construction pipeline**. We develop a four-stage pipeline that ensures subject consistency, diversity, and scalability. (iii) **Open-ended benchmark and evaluation protocol**. We introduce a benchmark with four sub-tasks and a rubricized VLM judge, and demonstrate that training on OpenSubject yields improvements on both our benchmark and external suites.

2 Related Work

Subject-driven image generation. Subject-driven image generation aims to synthesize identity-consistent images of specified subjects. Existing methods broadly fall into two architectural families: U-Net–based and DiT-based. Early approaches are predominantly U-Net–based; for example, DreamBooth [31] adapts Stable Diffusion [30] for single-subject personalization via test-time fine-tuning [7, 14, 32]. Zero-shot variants such as IP-Adapter [45], PhotoVerse [3], FastComposer [42], and FlashFace [50] improve inference efficiency by injecting identity features through vision encoders or lightweight adapters, yet they still inherit limitations of U-Net architectures. By contrast, DiT-based methods provide stronger representational capacity and more flexible conditioning. For single-subject generation, IC-LoRA [9], InstantID [35], and InfiniteYou [12] improve identity fidelity and prompt alignment via residual identity injection and attention-based control. Extending to multi-subject generation, methods including UNO [40], OminiControl [33], UniReal [4], DreamO [22], XVerse [2], OmniGen [39, 43], and Emu3.5 [5] explore multi-image conditioning, token-level constraints, and text-stream modulation to enable personalized control over multiple identities. Despite this progress, scalable and reliable multi-subject generation in open-ended scenes remains challenging, in large part due to the lack of a diverse and identity-consistent dataset. We address this gap with OpenSubject, a video-derived corpus for subject-driven generation and manipulation.

Datasets for subject-driven generation and manipulation. High-quality datasets are essential for preserving identity across diverse poses, illumination, and scenes, especially for multi-subject compositions. Prior work follows two paradigms: synthesis-based and retrieval-based. Synthesis-based corpora, such as Subjects200K [33] and UNO-1M [40], use strong base models (e.g., FLUX.1 [15]) and prompts to generate paired samples with controlled variation, but they may inherit generator biases and exhibit identity drift. Echo-4o-Image [46] further expands coverage with roughly 180K long-tailed samples, of which about 73K form paired multi-subject inputs. Retrieval-based datasets, exemplified by MultiID-2M [44], curate real-world group photos paired with identity references, but the availability of suitable multi-subject images limits them. OpenSubject

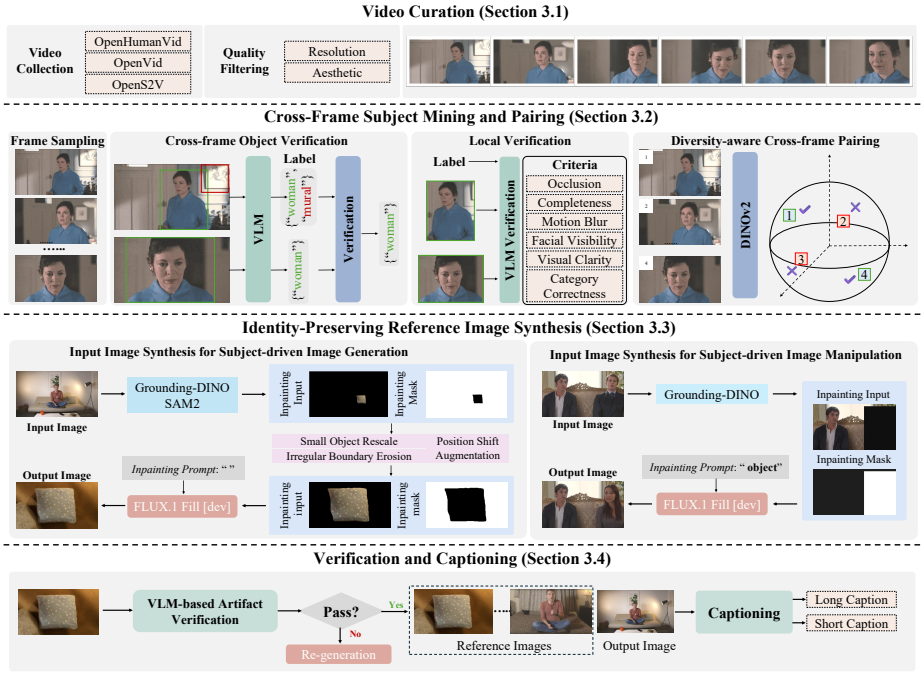

Fig. 2: Overview of the **OpenSubject** pipeline. (a) **Video curation**: collect open-source videos (OpenHumanVid, OpenVid, OpenS2V) and apply resolution/aesthetic filtering. (b) **Cross-frame subject mining and pairing**: enforce clip-level subject consensus with a vision–language model, localize instances with Grounding-DINO, and select diverse frame pairs. (c) **Identity-preserving reference synthesis**: synthesize training *inputs* via mask-guided outpainting (generation) and box-guided inpainting (manipulation). (d) **Verification and captioning**: perform VLM-based artifact checks with regeneration, then produce short and long captions for training.

complements these datasets by leveraging temporal consistency in videos to define instance-level identities, while synthesizing task-specific inputs to obtain scalable, diverse, and identity-consistent supervision for both subject-driven generation and manipulation.

3 OpenSubject

Fig. 2 provides an overview of the **OpenSubject** data-construction pipeline, which comprises video curation, cross-frame subject mining and pairing, identity-preserving reference image synthesis, and verification and captioning.

3.1 Video Curation

Video collection. We construct OpenSubject from large-scale, publicly available video corpora (OpenVid [23], OpenHumanVid [18], OpenS2V [49]).

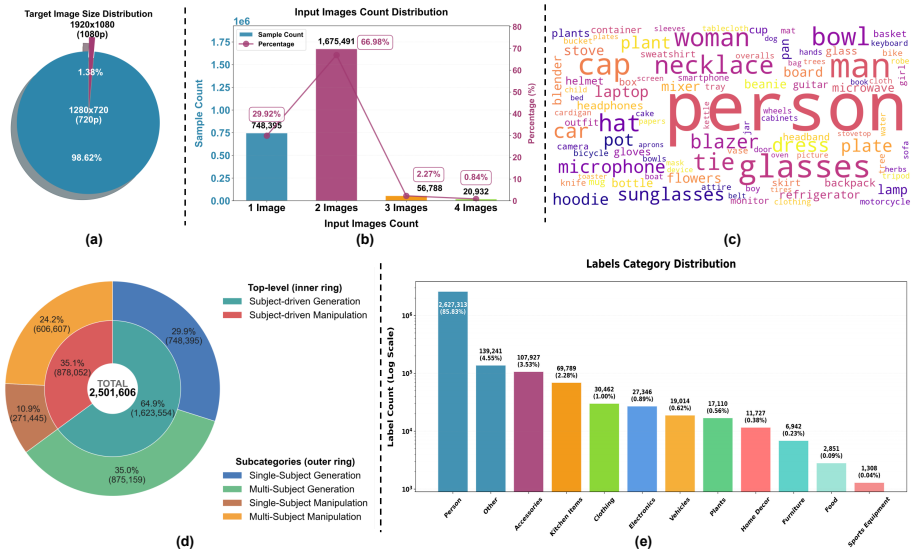

Fig. 3: Dataset statistics of OpenSubject. (a) Spatial resolution distributions. (b) Distribution of the number of references per sample. (c) Word cloud for subjects. (d) Task composition across four sub-tasks. (e) Subject category frequency.

Quality filtering. We retain only videos that satisfy minimum quality criteria, removing clips with spatial resolution below 720p or aesthetic scores below 5.8.

3.2 Cross-Frame Subject Mining and Pairing

To achieve identity consistency, our core strategy is to (i) segment videos into subject-persistent clips and (ii) enforce clip/instance validity through a strict verification stage (Sec. 3.4). We define a *subject-persistent clip* as a contiguous clip in which the target subject instance can be continuously tracked with stable category, without identity switch or long-term occlusion. Based on these reliable clips, we construct cross-frame training pairs using a hierarchical procedure: (i) uniformly sample four mid-range frames; (ii) enforce clip-level subject consensus, retaining only frames that share the same detected subject category; (iii) perform instance-level verification via grounding and a VLM to remove spurious detections; and (iv) select a diversity-aware pair by choosing the two frames with the largest DINOv2 [27] embedding distance.

Frame sampling. We uniformly sample four frames from the central temporal segment(s) after splitting each clip into five equal parts, avoiding the start/end segments, to capture natural appearance variation without redundancy, yielding candidates that differ in viewpoint and context.

Cross-frame subject verification. We apply Qwen2.5-VL-7B [1] to each candidate frame to detect foreground subjects and retain only instances occupying at least 5% of the image area. We then enforce cross-frame consensus by computing the intersection (or majority agreement) of subject categories across frames,

keeping only frames/instances whose category belongs to this consensus set. If the consensus set is empty, the clip is discarded.

Local verification. To ensure spatial and semantic validity, we perform instance-level grounding with Grounding-DINO [20] using class-specific confidence thresholds (0.5 for objects, 0.8 for humans). Detections are filtered with geometric priors—box count, relative area, aspect ratio, and pairwise IoU—to remove spurious or degenerate cases. The corresponding regions of interest are then cropped and evaluated by Qwen2.5-VL-7B [1] for (i) label correctness, (ii) occlusion, (iii) completeness, (iv) motion blur, and (v) facial visibility (for human subjects).

Diversity-aware cross-frame pairing. To broaden scene-context coverage, we compute DINOv2 embeddings for the filtered frames and select the pair with the largest cosine distance. Clips with fewer than two valid frames are discarded.

Role of DINOv2 in pairing. We emphasize that DINOv2 is used only for *appearance-diversity* selection, not for identity verification. Identity consistency is established *before* pairing through strict scene-level splits and cross-frame category consensus (this section); DINOv2 then merely selects the most visually diverse pair among frames already verified to share the same identity. As shown in Tab. 2(a), on a 5k-clip subset, choosing the maximum-distance (Max-DINOv2) pair markedly increases appearance diversity (LPIPS $0.31 \rightarrow 0.40$) with negligible change in identity similarity (ArcFace $0.619 \rightarrow 0.615$), confirming that diversity selection does not trade off identity.

3.3 Identity-Preserving Reference Image Synthesis

Given curated frames, we define two subject-conditioned tasks: subject-driven generation (inputs and targets drawn from different frames of the same clip) and subject-driven manipulation (edits within a single frame). Ground-truth targets are the original video frames, while inputs are synthesized via mask-guided outpainting (for generation) or inpainting (for manipulation). To standardize resolution and improve visual quality, we apply geometry-aware augmentations (scale and position jitter) and perform irregular boundary erosion on masks to reduce seam artifacts (e.g., banding or black borders) before synthesis.

Input image synthesis for subject-driven generation. Given verified subject labels and bounding boxes, we synthesize training inputs via mask-guided outpainting. (i) *Fine mask construction.* We refine localization with Grounding-DINO and SAM2 [29] to obtain instance masks. (ii) *Mask topology and geometry normalization.* To mitigate identity leakage, overlapping masks are reconciled by subtracting nested regions from larger instances. We then apply geometry augmentations (scale and placement): instances with area $< 30\%$ are upscaled to a random target in 30–40% (aspect ratio preserved), and all subjects are re-centered with bounded jitter to enhance layout diversity while maintaining identity fidelity. (iii) *Outpainting.* Using subject pixels and refined masks as constraints, we perform outpainting with FLUX.1 Fill [dev] [16], yielding identity-preserving inputs for subject-driven generation.

Input image synthesis for subject-driven manipulation. We construct input images for the manipulation task via mask-guided inpainting. (i) *Sample*

Table 2: Pipeline validation. **(a)** Cross-frame pairing on 5k clips: DINOv2 selects diverse appearance pairs (LPIPS↑) while identity (ArcFace) is largely unchanged. **(b)** Grounded verifier on a 500-sample holdout with 3-annotator majority labels (P/R/F1 vs. majority labels; Keep is the retention rate).

(a) Cross-frame pairing

Pairing	ArcFace ↑	LPIPS ↑
Random	0.619	0.31
Mid-quartile	0.622	0.35
Max-DINOv2 (ours)	0.615	**0.40**

(b) Grounded verifier

Rule	P	R	F1	Keep
LAION ≥ 4.0	0.62	0.93	0.74	87%
LAION ≥ 4.8	0.81	0.85	0.83	71%
LAION ≥ 5.5	0.93	0.42	0.58	24%
Qwen2.5-VL (ours)	**0.86**	**0.89**	**0.87**	67%

selection. We choose multi-object images with low pairwise box overlap and randomly select one or more target instances as editing regions. (ii) *Mask construction.* Unlike the generation branch, segmentation masks are not required. Instead, the bounding box is used to define the erase mask and the corresponding input conditioning. (iii) *Inpainting.* We apply FLUX.1 Fill [dev] to complete the erased regions. The resulting pairs provide precise localization supervision for referent replacement while preserving non-target content.

3.4 Verification and Captioning

We adopt a verify–refine–caption stage. Qwen2.5-VL-7B acts as the verifier, assessing synthesized samples for artifacts and physical plausibility. Failed samples are re-synthesized with a different random seed. For accepted samples, Qwen2.5-VL-7B generates two caption variants (short and long), with the instruction style randomly instantiated as either generation or editing.

Reliability of the verifier. Our verifier is not an unvalidated global VLM filter: candidates are first localized by Grounding-DINO and then checked by the VLM for label correctness, occlusion, blur, and facial visibility. We validate it on a 500-sample pre-filtering holdout with 3-annotator majority labels, covering both likely-valid samples and hard failures (e.g., blur, occlusion, missing limbs, incomplete faces). As shown in Tab. 2(b), our Grounding-DINO → Qwen2.5-VL verifier achieves the best F1 (0.87) with balanced precision/recall at a 67% keep rate, whereas LAION-aesthetic thresholds either over-filter or admit invalid pairs.

3.5 Statistical Analysis

We construct the **OpenSubject** dataset with the above pipeline, yielding 2.5M samples and 4.35M images. As shown in Fig. 3 (a–e), resolutions are predominantly 720p, with a smaller 1080p subset. The reference set includes a single-image subset (748k), while the remaining samples contain two or more input images, enabling both single- and multi-reference settings. The corpus spans four sub-tasks: single-subject generation, single-subject manipulation, multi-subject generation, and multi-subject manipulation. Category and scene coverage is broad, encompassing people, objects, and diverse environments.

4 Benchmark

Prior benchmarks (e.g., XVerse [2]) focus on clean, single-subject portraits and rarely evaluate subjects in complex scenes; moreover, subject-driven manipulation is typically omitted. We introduce a benchmark (named OSBench) that spans subject-driven generation and manipulation tasks.

Tasks. The benchmark comprises four sub-tasks (150 samples each): (i) **Single-subject generation**. Synthesize an identity-consistent image from one reference under an open-ended prompt; (ii) **Multi-subject generation**. Synthesize an image by fusing two to four references under an open-ended prompt; (iii) **Single-subject manipulation**. Replace one target in a scene that contains a single principal object; and (iv) **Multi-subject manipulation**. Replace one target in a complex scene (containing multiple subjects) while preserving non-target content. OSBench is constructed from videos disjoint from the OpenSubject training split.

Evaluation protocol. Following instruction-based assessment methods (e.g., VIEScore [13], OmniContext [39]), we use a strong VLM judge (GPT-4.1 [24]) to assign 0–10 scores with rubricized prompts and independent criteria, and report the average score over all samples. We complement VLM-based protocol with human assessment (Sec. 5.2). For **generation** tasks, we measure (i) **Prompt Adherence (PA)**, assessing compliance with the prompt in terms of attributes, counts, and relations; and (ii) **Identity Fidelity (IF)**, assessing whether the generated subject matches the provided reference identity across views. We report **Overall** as the geometric mean of PA and IF. For **manipulation** tasks, we measure (i) **Manipulation Fidelity (MF)**, assessing whether the edited region matches the referenced subject(s) and requested change; and (ii) **Background Consistency (BC)**, assessing whether non-edited regions remain unchanged. We report **Overall** as the sample-level geometric mean of MF and BC, which favors methods that are simultaneously strong on both dimensions.

5 Experiments

5.1 Experimental Setup

Evaluation baselines. We evaluate both closed- and open-source models. Closed-source methods include Gemini 2.5 Flash Image Preview [34], GPT-4o [25], and Gemini 2.0 Flash Image Preview [8]. Open-source subject-driven models include UNO [40], DreamO [22], XVerse [2], DreamOmni2 [41], Qwen-Image-Edit-2509 [38], and OmniGen2 [39]. In addition, to quantify the utility of OpenSubject beyond prior corpora, we report controlled fine-tuning experiments on a fixed base model (OmniGen2) under matched compute budgets.

Implementation details. For closed-source models, we use the official inference APIs; when an API does not expose output-resolution control, we specify the target resolution in the prompt. For open-source models, we use the official checkpoints and default inference settings. Because Gemini 2.5 Flash Image Preview, Gemini 2.0 Flash Image Preview, and GPT-4o do not permit setting the output resolution, we request the desired resolution via the prompt. All

Table 3: Quantitative results of single-subject and multi-subject driven generation and manipulation on the OSBench.

Method	Subject-driven generation						Subject-driven manipulation						Average ↑
	Single-Subject			Multi-Subject			Single-Subject			Multi-Subject			
	PA	IF	Overall	PA	IF	Overall	MF	BC	Overall	MF	BC	Overall	
Closed-source models													
Gemini 2.5 Flash Image Preview [34]	9.05	8.80	8.90	8.18	8.40	8.22	7.25	8.18	7.22	6.88	4.90	5.15	7.37
GPT-4o-2024-11-20 [25]	9.20	8.52	8.83	8.37	7.60	7.80	7.62	6.25	6.78	7.08	2.95	3.76	6.79
Gemini 2.0 Flash Image Preview [8]	8.35	8.28	8.29	7.60	7.40	7.41	3.40	3.18	3.22	5.00	2.40	2.80	5.43
Open-source models													
UNO [40]	7.55	5.28	5.93	7.12	6.10	6.38	2.80	0.56	0.80	1.46	0.86	0.90	3.50
DreamO [22]	7.12	4.10	5.10	7.05	6.15	6.50	3.82	0.29	0.40	2.10	0.72	0.74	3.19
XVerse [2]	6.60	1.92	3.05	7.02	5.25	5.92	3.12	0.40	0.62	0.84	0.69	0.70	2.57
DreamOmni2 [41]	8.02	7.32	7.50	7.44	6.60	6.85	6.05	6.28	5.70	5.08	1.42	2.18	5.56
Qwen-Image-Edit-2509 [38]	9.23	8.31	8.72	8.33	6.96	7.55	7.72	6.41	6.73	7.25	3.93	4.82	6.95
OmniGen2 [39]	8.91	8.39	8.61	8.20	7.68	7.88	5.93	5.13	5.08	7.35	3.39	4.51	6.52

other models generate one image per item at a fixed resolution. For OSBench, we evaluate four sub-tasks (single-/multi-subject generation and single-/multi-subject manipulation), with 150 samples per sub-task (600 in total), and report both sub-task metrics and the overall average score.

Training data and protocol for fine-tuned models. We consider the following training sets. (i) **OpenSubject**: 500k samples randomly sampled from OpenSubject dataset. (ii) **T2I**: 100k internal text-to-image samples used to preserve prompt-following ability. We train models on 16× H800 GPUs.

5.2 Comparison Experiments

OSBench Evaluation. As shown in Tab. 3, most approaches struggle on *multi-subject generation* and *subject-driven manipulation*. Open-source models such as UNO, DreamO, and XVerse exhibit very low manipulation performance, indicating weak identity replacement and preservation in complex scenes. Qwen-Image-Edit-2509 is the strongest open-source baseline for manipulation but remains limited in multi-subject cases (Overall 4.82). Closed-source models also drop markedly from generation to manipulation. For example, Gemini 2.5 Flash Image Preview attains only 5.15 Overall on multi-subject manipulation despite solid generation scores. In summary, current models are not robust when multiple identities must be fused or replaced while keeping non-target content intact.

Effectiveness of the OpenSubject Dataset. We assess OpenSubject by fine-tuning two representative base models, *OmniGen2* and *Qwen-Image-Edit-2509*. For each model, we augment training with synthetic T2I data and sampled OpenSubject instances, then fine-tune under identical settings. As shown in Tab. 4, OpenSubject consistently improves both models, with the largest gains in manipulation: OmniGen2 improves from 6.52 to 7.12 (+0.6) on Average and boosts multi-subject manipulation Overall by +1.75 (4.51→6.26), while Qwen-Image-Edit-2509 increases Average from 6.95 to 8.18 (+1.23) and raises multi-subject manipulation Overall by +2.34 (4.82→7.16). Improvements also appear in multi-subject generation, indicating stronger multi-identity composition.

Table 4: Evaluation results of fine-tuning with OpenSubject. Rows with a light cyan background denote models fine-tuned with OpenSubject; the Δ row reports the score differences w.r.t. the corresponding baseline. Green and red indicate performance improvements and decreases, respectively.

Method	Subject-driven generation						Subject-driven manipulation						Average↑
	Single-Subject			Multi-Subject			Single-Subject			Multi-Subject			
	PA	IF	Overall	PA	IF	Overall	MF	BC	Overall	MF	BC	Overall	
OmniGen2	8.91	8.39	8.61	8.20	7.68	7.88	5.93	5.13	5.08	7.35	3.39	4.51	6.52
+ Ours	8.78	8.80	8.67	8.35	8.34	8.15	6.53	5.23	5.41	7.65	5.49	6.26	7.12
Δ	(-0.13)	(+0.41)	(+0.06)	(+0.15)	(+0.66)	(+0.27)	(+0.60)	(+0.10)	(+0.33)	(+0.30)	(+2.10)	(+1.75)	(+0.60)
Qwen-Image-Edit-2509	9.23	8.31	8.72	8.33	6.96	7.55	7.72	6.41	6.73	7.25	3.93	4.82	6.95
+ Ours	9.05	9.03	9.02	8.59	8.63	8.57	7.38	8.96	7.96	7.46	8.43	7.16	8.18
Δ	(-0.18)	(+0.72)	(+0.30)	(+0.26)	(+1.67)	(+1.02)	(-0.34)	(+2.55)	(+1.23)	(+0.21)	(+4.50)	(+2.34)	(+1.23)

Table 5: Ablation study of fine-tuning with OpenSubject. Green and red indicate performance improvements and decreases, respectively.

Method	Subject-driven generation						Subject-driven manipulation						Average↑
	Single-Subject			Multi-Subject			Single-Subject			Multi-Subject			
	PA	IF	Overall	PA	IF	Overall	MF	BC	Overall	MF	BC	Overall	
OmniGen2 (baseline)	8.91	8.39	8.61	8.20	7.68	7.88	5.93	5.13	5.08	7.35	3.39	4.51	6.52
+ T2I	9.07	7.86	8.35	8.35	7.41	7.73	5.07	3.19	3.53	6.42	1.64	3.01	5.66
Δ	(+0.16)	(-0.53)	(-0.26)	(+0.15)	(-0.27)	(-0.15)	(-0.86)	(-1.94)	(-1.55)	(-0.93)	(-1.75)	(-1.50)	(-0.86)
+ **OpenSubject (ours)**	8.78	8.80	8.67	8.35	8.34	8.15	6.53	5.23	5.41	7.65	5.49	6.26	7.12
Δ	(-0.13)	(+0.41)	(+0.06)	(+0.15)	(+0.66)	(+0.27)	(+0.60)	(+0.10)	(+0.33)	(+0.30)	(+2.10)	(+1.75)	(+0.60)

Human evaluation. Following prior work, we use a VLM [39] as a scalable judge, and we additionally perform a human study (Tab. 8) with six annotators based on our benchmark. Human scores follow the same trend as VLM-based results, and fine-tuned models remain preferred over the baseline, supporting the main conclusion that OpenSubject can improve subject-driven generation and manipulation performance.

5.3 Ablation Experiments

Dataset Ablations. We evaluate three OmniGen2 configurations: the baseline, the baseline fine-tuned with T2I only, and fine-tuned on OpenSubject plus T2I. As shown in Tab. 5, adding only T2I yields gains in prompt adherence for generation but reduces identity fidelity and weakens manipulation. Fine-tuning on OpenSubject reverses these declines and produces consistent improvements across tasks, raising the "Average" from 6.52 to 7.12.

Dataset Comparison. Tab. 7 compares different training data sources under the same base model and training protocol. We report each prior dataset at its native released scale (Echo-4o-Image: 73k, MICo: 150k, UNO: 1M), and additionally report OpenSubject at 73k/150k/500k to expose scaling behavior under the same recipe. Overall, **T2I + Ours (500k)** achieves the best Average score (7.12), improving over the baseline (6.52) by +0.60 and outperforming **T2I + Echo-4o-Image** (6.12), **T2I + UNO** (5.52), and **T2I + MICo-150k** (6.02). From a task-level view, OpenSubject performs best on the most challenging composition and manipulation settings: it reaches the best multi-subject generation Overall (8.15, 500k), the best multi-subject generation PA

Table 6: Pipeline ablations at 50k (OmniGen2, identical LoRA recipe). Each row removes one component. Values are per sub-task VLM Overall and the Macro Avg.

Configuration	Single-Subj. (Gen)	Multi-Subj. (Gen)	Single-Subj. (Manip)	Multi-Subj. (Manip)	Macro Avg.
Ours (50k)	**8.44**	**7.94**	**5.21**	**5.45**	**6.76**
w/o cross-frame VLM consensus	8.26	7.70	4.97	5.19	6.53
w/o Grounding-DINO local verify	8.28	7.74	4.99	5.23	6.56
w/o verify-refine loop	8.32	7.80	5.05	5.27	6.61
w/o irregular boundary erosion	8.34	7.82	5.07	5.29	6.63
w/o geometry-aware aug.	8.36	7.84	5.11	5.33	6.66

(8.35, 500k) and IF (8.34, 500k), the best single-subject manipulation Overall (5.48, 150k), and the best multi-subject manipulation Overall (6.26, 500k). It also achieves the highest multi-subject manipulation MF/BC (7.65/5.49, 500k). In addition, OpenSubject yields the strongest single-subject manipulation MF (6.64, 150k) and the best multi-subject manipulation MF (7.65, 500k), highlighting its robustness for harder edits. Compared with prior datasets, Echo-4o-Image mainly improves prompt adherence in single-subject generation (PA 8.95) but is weaker on manipulation metrics (single-/multi-subject manipulation Overall: 4.30/3.86); UNO shows limited gains and the lowest Average score (5.52); MICo-150k improves background consistency in single-subject manipulation (BC 5.77) but underperforms on multi-subject manipulation (Overall 3.07). Overall, these results suggest that OpenSubject provides a better balance between prompt following, identity fidelity, and manipulation robustness.

Pipeline component ablations. To isolate the contribution of each pipeline stage, we rebuild a 50k subset five times, each time removing a single component, and fine-tune OmniGen2 under the same LoRA recipe. As shown in Tab. 6, removing any component reduces the VLM macro-average; cross-frame VLM consensus and Grounding-DINO local verification contribute the most, confirming that identity-grounded mining and verification are the key drivers of data quality.

Face Consistency. We do not use an explicit face-recognition loss in the pipeline. Identity consistency is enforced by strict video-level splits (detection and tracking for validation and filtering), so samples within each clip are inherently consistent. To quantify identity quality, we measured ArcFace similarity, obtaining an average of 0.6059. We further manually verified the 500 pairs (absolute bottom-500) with the lowest similarity and confirmed they still correspond to the same identity, with lower scores mainly due to pose and camera-angle changes.

5.4 Quantitative Evaluation on Other Benchmarks

OmniContext [39]. As shown in Tab. 9, fine-tuning OmniGen2 on OpenSubject improves the overall average score from 7.18 to 7.34 (+0.16). The gains are most pronounced in settings that require integrating multiple references and grounding them in complex scenes: *MULTIPLE/Character* improves from 7.11 to 7.34 (+0.23), *MULTIPLE/Object* from 7.13 to 7.37 (+0.24), and *MULTIPLE/Char.+Obj.* from 7.45 to 7.87 (+0.42). Scene-level categories also improve

Table 7: Dataset comparison and data-scaling study on OmniGen2. We compare prior datasets (Echo-4o-Image, UNO, MICo-150k) at their native scales and OpenSubject at different sample sizes, using the same training recipe. Bold and underlined numbers indicate the best and second-best results in each column, respectively.

Method	Subject-driven generation						Subject-driven manipulation						Average↑
	Single-Subject			Multi-Subject			Single-Subject			Multi-Subject			
	PA	IF	Overall	PA	IF	Overall	MF	BC	Overall	MF	BC	Overall	
OmniGen2 (baseline)	8.91	8.39	8.61	8.20	7.68	7.88	5.93	5.13	5.08	7.35	3.39	4.51	6.52
+ **Echo-4o-Image (73k)**	**8.95**	8.43	<u>8.66</u>	7.92	7.48	7.63	5.52	3.95	4.3	7.12	2.53	3.86	6.12
+ **UNO (1M)**	8.70	8.32	8.48	7.88	7.60	7.62	5.32	3.00	3.40	5.65	1.50	2.57	5.52
+ **MICo-150k (150k)**	8.58	8.18	8.34	<u>8.03</u>	7.42	7.63	5.68	**5.77**	5.03	6.82	1.97	3.07	6.02
+ **Ours (73k)**	8.36	8.63	8.46	7.92	8.10	7.95	6.43	5.00	5.32	7.38	4.20	5.62	6.84
+ **Ours (150k)**	8.39	<u>8.70</u>	8.49	7.95	<u>8.20</u>	<u>8.07</u>	**6.64**	5.14	**5.48**	<u>7.54</u>	<u>4.44</u>	<u>5.88</u>	<u>6.98</u>
+ **Ours (500k)**	<u>8.78</u>	**8.80**	**8.67**	**8.35**	**8.34**	**8.15**	<u>6.53</u>	<u>5.23</u>	<u>5.41</u>	**7.65**	**5.49**	**6.26**	**7.12**

Table 8: Human study results of fine-tuning with OpenSubject. Scores are rated on a 1–7 scale (higher is better), averaged across six annotators.

Method	Subject-driven generation				Subject-driven manipulation			
	Single-Subject		Multi-Subject		Single-Subject		Multi-Subject	
	PA	IF	PA	IF	MF	BC	MF	BC
Qwen-Image-Edit-2509	6.10	5.62	5.14	4.76	5.05	4.29	4.02	3.08
Qwen-Image-Edit-2509 + Ours	6.12	6.18	6.08	5.66	4.92	6.02	6.05	6.41

(*SCENE/Character*: 6.38→6.50; *SCENE/Object*: 6.71→6.92), whereas the single-reference case *SINGLE/Object* remains nearly unchanged (7.58→7.54). Overall, these results suggest that OpenSubject primarily strengthens identity reasoning and compositional control under multi-entity and context-rich settings.

ImgEdit [47]. As shown in Tab. 10, fine-tuning OmniGen2 on **OpenSubject** improves instruction-based editing, increasing the overall score from 3.44 to 3.79 (+0.35). The largest gains are observed in categories that require precise localization or structural modifications: *Extract* improves from 1.77 to 2.61 (+0.84), *Hybrid* from 2.52 to 3.28 (+0.76), *Add* from 3.57 to 4.28 (+0.71), and *Background* from 3.57 to 4.13 (+0.56). Moderate improvements are found for *Replace* and *Remove* (+0.23 each), as well as *Adjust* (+0.21), while *Style* (4.81→4.66) and *Action* (4.68→4.45) slightly decrease. Overall, the results suggest that OpenSubject fine-tuning enhances edit performance.

5.5 Qualitative Results

Fig. 4 presents comparisons across methods, with colored dashed boxes indicating corresponding regions of interest. Here, "Ours" denotes OmniGen2 fine-tuned on OpenSubject. Many models struggle to preserve identity in multi-subject scenes and often introduce changes outside the intended edits. Compared with the original OmniGen2, our fine-tuned model better maintains subject identity, follows attribute specifications, and produces more coherent multi-subject compositions.

Table 9: Quantitative comparison results on OmniContext. "Char. + Obj." indicates Character + Object. Fine-tuning on **OpenSubject** yields large gains on the MULTIPLE subset and consistent improvements on SINGLE and SCENE.

Model	Size (B)	SINGLE		MULTIPLE			SCENE			Average↑
		Character	Object	Character	Object	Char. + Obj.	Character	Object	Char. + Obj.	
Closed-source models										
FLUX.1 Kontext [Max] [17]	-	8.48	8.68	-	-	-	-	-	-	-
Gemini 2.5 Flash Image Preview [34]	-	8.52	9.14	7.80	8.64	6.63	6.74	7.11	6.04	7.58
Gemini 2.5 Flash Image [34]	-	8.62	8.91	7.88	8.92	7.39	7.29	7.05	6.68	7.84
GPT-4o [25]	-	8.90	9.01	9.07	8.95	8.54	8.90	8.44	8.60	8.80
Gemini 2.0 Flash [8]	-	5.06	5.17	2.91	2.16	3.80	3.02	3.89	2.92	3.62
Open-source models										
Emu3.5 [5]	32	8.72	9.46	8.65	9.09	8.78	8.78	8.89	8.15	8.82
Qwen-Image-Edit-2509 [38]	20	8.35	9.13	7.65	8.85	7.90	5.16	7.75	6.73	7.69
OmniGen [43]	3.8	7.21	5.71	5.65	5.44	4.68	3.59	4.32	5.12	4.34
InfiniteYou [12]	12	6.05	-	-	-	-	-	-	-	-
UNO [40]	12	6.60	6.83	2.54	6.51	4.39	2.06	4.33	4.37	4.71
BAGEL [6]	14	5.48	7.03	5.17	6.64	6.24	4.07	5.71	5.47	5.73
Baseline (OmniGen2 [39])	7	8.05	7.58	7.11	7.13	7.45	6.38	6.71	7.04	7.18
Ours	7	8.18 (+0.13)	7.54 (-0.04)	7.34 (+0.23)	7.37 (+0.24)	7.87 (+0.42)	6.50 (+0.12)	6.92 (+0.21)	7.00 (-0.04)	7.34 (+0.16)

Table 10: Quantitative results on ImgEdit benchmark [47]. Fine-tuning on **OpenSubject** also yields gains in image editing task, especially on the "Add", "Extract", "Background", and "Hybrid" subsets. The fine-tuned omnigen2 model achieves a higher overall score than FLUX.1 Kontext [Dev] [17].

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
Closed-source models										
FLUX.1 Kontext [Pro] [17]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT-Image-1 [High] [26]	4.61	4.33	2.90	4.35	3.66	**4.57**	**4.93**	**3.96**	**4.89**	4.20
Gemini 2.5 Flash Image Preview [34]	4.47	4.19	3.81	4.39	**4.70**	4.20	4.18	3.48	4.68	4.23
Gemini 2.5 Flash Image [34]	**4.65**	4.34	3.69	4.49	4.65	4.32	4.13	3.66	4.59	4.28
Open-source models										
AnyEdit [48]	3.18	2.95	1.88	2.47	2.23	2.24	2.85	1.56	2.65	2.45
UltraEdit [52]	3.44	2.81	2.13	2.96	1.45	2.83	3.76	1.91	2.98	2.70
OmniGen [43]	3.47	3.04	1.71	2.94	2.43	3.21	4.19	2.24	3.38	2.96
ICEdit [51]	3.58	3.39	1.73	3.15	2.93	3.08	3.84	2.04	3.68	3.05
Step1X-Edit [21]	3.88	3.14	1.76	3.40	2.41	3.16	4.63	2.64	2.52	3.06
BAGEL [6]	3.56	3.31	1.70	3.3	2.62	3.24	4.49	2.38	4.17	3.20
UniWorld-1 [19]	3.82	3.64	2.27	3.47	3.24	2.99	4.21	2.96	2.74	3.26
Lego-Edit [11]	3.67	3.82	2.47	3.22	3.39	4.47	4.01	3.18	3.24	3.50
FLUX.1 Kontext [Dev] [17]	4.12	3.80	2.04	4.22	3.09	3.97	4.51	3.35	4.25	3.71
Qwen-Image-Edit [38]	4.38	4.16	3.43	4.66	4.14	4.38	4.81	3.82	4.69	4.27
Qwen-Image-Edit-2509 [38]	4.32	**4.36**	**4.04**	4.64	4.52	4.37	4.84	3.39	4.71	4.35
Emu3.5 [5]	4.61	4.32	3.96	**4.84**	4.58	4.35	4.79	3.69	4.57	**4.41**
Baseline (OmniGen2 [39])	3.57	3.06	1.77	3.74	3.20	3.57	4.81	2.52	4.68	3.44
Ours	4.28 (+0.71)	3.27 (+0.21)	2.61 (+0.84)	3.97 (+0.23)	3.43 (+0.23)	4.13 (+0.56)	4.66 (-0.15)	3.28 (+0.76)	4.45 (-0.23)	3.79 (+0.35)

For manipulation, edits remain confined to the required regions while non-target content is preserved, whereas other methods often show identity drift or unintended alterations beyond the edit area.

5.6 Limitations and Discussion

Video-frame target domain gap. Although OpenSubject uses real video frames as supervision targets, these frames can still contain compression artifacts and motion noise. Empirically, we do not observe an obvious "video look" in generated results, and gains on OmniContext and ImgEdit suggest this gap is not dominant in our current setting. We attribute this to three factors: (i)

Fig. 4: Qualitative comparison. Colored dashed boxes mark regions of interest for comparison. Boxes of the same color denote corresponding regions across methods and refer to the related area in the input image.

multi-stage filtering removes many low-quality frames; (ii) diverse synthesized inputs discourage memorization of video-specific low-level patterns; and (iii) fine-tuning starts from strong image priors and jointly uses T2I regularization.

Synthetic-input bias. Our input construction relies on FLUX.1 Fill to synthesize inpainting/outpainting contexts. This improves scalability but can introduce generator-specific priors into training inputs.

6 Conclusion

This paper addresses a central bottleneck in subject-driven image generation and manipulation: the lack of large-scale data that jointly supports subject consistency, compositional diversity, and controllable editing. We introduce **OpenSubject**, a video-derived dataset (2.5M samples, 4.35M images) built to provide scalable subject-consistent supervision for both generation and manipulation. We also establish an open-ended benchmark covering single- and multi-subject settings with aspect-specific evaluation criteria. Experiments on our benchmark and external suites show that existing models degrade substantially in complex multi-subject scenarios, while fine-tuning with **OpenSubject** consistently improves identity fidelity, compositional control, and manipulation robustness.

References

[1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)

[2] Chen, B., Zhao, M., Sun, H., Chen, L., Wang, X., Du, K., Wu, X.: XVerse: Consistent multi-subject control of identity and semantic attributes via DiT modulation. arXiv preprint arXiv:2506.21416 (2025)

[3] Chen, L., Zhao, M., Liu, Y., Ding, M., Song, Y., Wang, S., Wang, X., Yang, H., Liu, J., Du, K., et al.: PhotoVerse: Tuning-free image customization with text-to-image diffusion models. arXiv preprint arXiv:2309.05793 (2023)

[4] Chen, X., Zhang, Z., Zhang, H., Zhou, Y., Kim, S.Y., Liu, Q., Li, Y., Zhang, J., Zhao, N., Wang, Y., et al.: UniReal: Universal image generation and editing via learning real-world dynamics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12501–12511 (2025)

[5] Cui, Y., Chen, H., Deng, H., Huang, X., Li, X., Liu, J., Liu, Y., Luo, Z., Wang, J., Wang, W., et al.: Emu3.5: Native multimodal models are world learners. arXiv preprint arXiv:2510.26583 (2025)

[6] Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)

[7] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)

[8] Google: Gemini 2.0 Flash. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation (2025), accessed on 2025-11-10

[9] Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-Context LoRA for diffusion transformers. arXiv preprint arXiv:2410.23775 (2024)

[10] Huang, Q., Fu, S., Liu, J., Jiang, H., Yu, Y., Song, J.: Resolving multi-condition confusion for finetuning-free personalized image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3707–3714 (2025)

[11] Jia, Q., Liu, Y., Chai, Y., Yao, X., Lu, Q., Zhang, Y., Shi, R., Huang, Y., Zhang, G.: Lego-Edit: A general image editing framework with model-level bricks and mllm builder. arXiv preprint arXiv:2509.12883 (2025)

[12] Jiang, L., Yan, Q., Jia, Y., Liu, Z., Kang, H., Lu, X.: InfiniteYou: Flexible photo recrafting while preserving your identity. arXiv preprint arXiv:2503.16418 (2025)

[13] Ku, M., Jiang, D., Wei, C., Yue, X., Chen, W.: VIEScore: Towards explainable metrics for conditional image synthesis evaluation. arXiv preprint arXiv:2312.14867 (2023)

[14] Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)

[15] Labs, B.F.: FLUX. https://github.com/black-forest-labs/flux (2024), accessed on 2025-11-08
[16] Labs, B.F.: FLUX.1-Fill-dev. https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev (2025), hugging Face model card, accessed on 2025-11-09
[17] Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al.: FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
[18] Li, H., Xu, M., Zhan, Y., Mu, S., Li, J., Cheng, K., Chen, Y., Chen, T., Ye, M., Wang, J., et al.: OpenHumanVid: A large-scale high-quality dataset for enhancing human-centric video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7752–7762 (2025)
[19] Lin, B., Li, Z., Cheng, X., Niu, Y., Ye, Y., He, X., Yuan, S., Yu, W., Wang, S., Ge, Y., et al.: UniWorld-V1: High-resolution semantic encoders for unified visual understanding and generation. arXiv preprint arXiv:2506.03147 (2025)
[20] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
[21] Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al.: Step1X-Edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025)
[22] Mou, C., Wu, Y., Wu, W., Guo, Z., Zhang, P., Cheng, Y., Luo, Y., Ding, F., Zhang, S., Li, X., et al.: DreamO: A unified framework for image customization. arXiv preprint arXiv:2504.16915 (2025)
[23] Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
[24] OpenAI: GPT-4.1. https://openai.com/index/gpt-4-1 (2025), accessed on 2025-11-12
[25] OpenAI: GPT-4o. https://openai.com/index/introducing-4o-image-generation (2025), accessed on 2025-11-11
[26] OpenAI: Image generation API. https://openai.com/index/image-generation-api/ (2025), accessed on 2025-11-14
[27] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
[28] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
[29] Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

[30] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
[31] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
[32] Ruiz, N., Li, Y., Jampani, V., Wei, W., Hou, T., Pritch, Y., Wadhwa, N., Rubinstein, M., Aberman, K.: HyperDreamBooth: Hypernetworks for fast personalization of text-to-image models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6527–6536 (2024)
[33] Tan, Z., Liu, S., Yang, X., Xue, Q., Wang, X.: OminiControl: Minimal and universal control for diffusion transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14940–14950 (2025)
[34] Team, G.: Gemini 2.5 Flash & Gemini 2.5 Flash Image Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf (2025), accessed on 2025-11-16
[35] Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A., Li, H., Tang, X., Hu, Y.: InstantID: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
[36] Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: MS-Diffusion: Multi-subject zero-shot image personalization with layout guidance. arXiv preprint arXiv:2406.07209 (2024)
[37] Wei, X., Cen, K., Wei, H., Guo, Z., Cui, K., Li, B., Wang, Z., Zhang, J., Zhang, L.: MICo-150K: A comprehensive dataset advancing multi-image composition. arXiv preprint arXiv:2512.07348 (2025)
[38] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-Image technical report. arXiv preprint arXiv:2508.02324 (2025)
[39] Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., et al.: OmniGen2: Towards instruction-aligned multimodal generation. arXiv preprint arXiv:2506.18871 (2025)
[40] Wu, S., Huang, M., Wu, W., Cheng, Y., Ding, F., He, Q.: Less-to-more generalization: Unlocking more controllability by in-context generation. arXiv preprint arXiv:2504.02160 (2025)
[41] Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., et al.: DreamOmni2: Multimodal instruction-based editing and generation. arXiv preprint arXiv:2510.06679 (2025)
[42] Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: FastComposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision **133**(3), 1175–1194 (2025)
[43] Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: OmniGen: Unified image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13294–13304 (2025)

[44] Xu, H., Cheng, W., Xing, P., Fang, Y., Wu, S., Wang, R., Zeng, X., Jiang, D., Yu, G., Ma, X., et al.: WithAnyone: Towards controllable and ID consistent image generation. arXiv preprint arXiv:2510.14975 (2025)
[45] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
[46] Ye, J., Jiang, D., Wang, Z., Zhu, L., Hu, Z., Huang, Z., He, J., Yan, Z., Yu, J., Li, H., et al.: Echo-4o: Harnessing the power of GPT-4o synthetic images for improved image generation. arXiv preprint arXiv:2508.09987 (2025)
[47] Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: ImgEdit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275 (2025)
[48] Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: AnyEdit: Mastering unified high-quality image editing for any idea. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26125–26135 (2025)
[49] Yuan, S., He, X., Deng, Y., Ye, Y., Huang, J., Lin, B., Luo, J., Yuan, L.: OpenS2V-Nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. arXiv preprint arXiv:2505.20292 (2025)
[50] Zhang, S., Huang, L., Chen, X., Zhang, Y., Wu, Z.F., Feng, Y., Wang, W., Shen, Y., Liu, Y., Luo, P.: FlashFace: Human image personalization with high-fidelity identity preservation. arXiv preprint arXiv:2403.17008 (2024)
[51] Zhang, Z., Xie, J., Lu, Y., Yang, Z., Yang, Y.: In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. arXiv preprint arXiv:2504.20690 (2025)
[52] Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: UltraEdit: Instruction-based fine-grained image editing at scale. Advances in Neural Information Processing Systems **37**, 3058–3093 (2024)