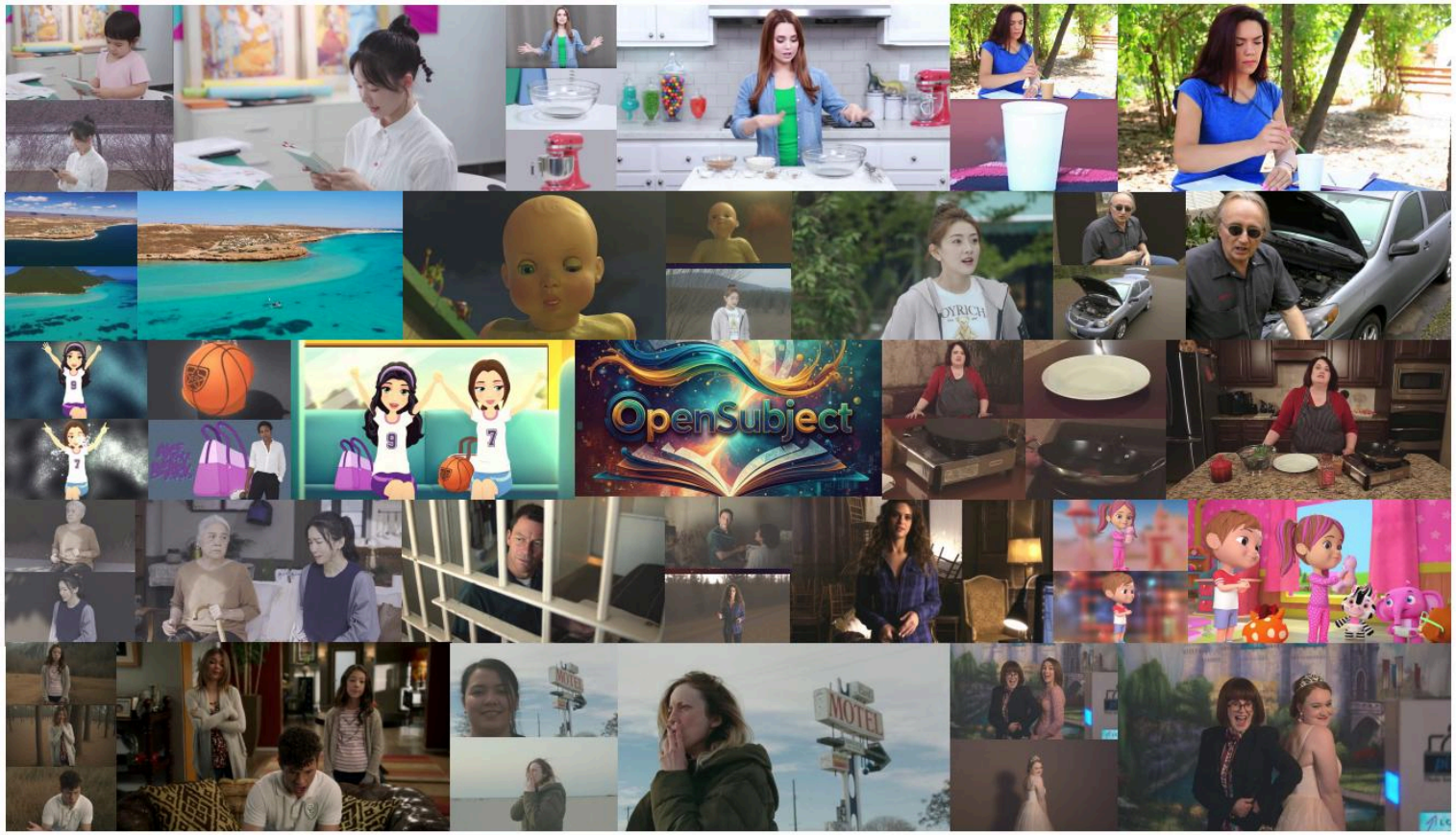
OpenSubject: Leveraging Video-Derived Identity and Diversity Priors for Subject-driven Image Generation and Manipulation

Yexin Liu, Manyuan Zhang, Yueze Wang, Hongyu Li, Dian Zheng, Weiming Zhang, Changsheng Lu, Yan Feng, Peng Pei, Xunliang Cai, Harry Yang

HKUST, Meituan, Independent Researcher, HKUST(GZ)

The Challenge in Subject-Driven Generation

- Despite promising progress, current models often struggle to preserve reference identities, especially in complex scenes with multiple subjects.
- Existing dataset creation methods have significant limitations.



Our Solution: The OpenSubject Corpus

- We introduce OpenSubject, a large-scale corpus with 2.5M samples derived from video data to provide identity-consistent supervision.
- Videos naturally offer multi-frame observations of subjects with rich variations in viewpoint, illumination, and environment.
- Our approach addresses key challenges in dataset construction: maximizing diversity while preserving identity and ensuring sufficient context variation.

Comparison with Existing Datasets

- OpenSubject is the first large-scale dataset to leverage video as a source of identity priors.
- It uniquely supports subject-driven manipulation and provides extensive multi-subject samples with diverse contexts for general objects.

Dataset	Paired Single-subject Input Samples	Paired Multi-subject Input Samples	Subject-driven Manipulation	IP Source	General Objects	Diverse Contexts
Echo-4o-Image [47]	✗	73k	✗	Synthesis	✓	✗
UNO-1M [40]	1M	✗	✗	Synthesis	✓	✗
Subjects200K [44]	200k	✗	✗	Synthesis	✓	✗
MultiID-2M [45]	✗	500k	✗	Retrieval	✗	✗
OpenSubject (ours)	748k	1752k	✓	Video	✓	✓

The OpenSubject Construction Pipeline

01 **1. Video Curation**

Collect videos from large-scale public corpora and apply resolution (>720p) and aesthetic filters to ensure high quality.

02 **2. Cross-Frame Mining & Pairing**

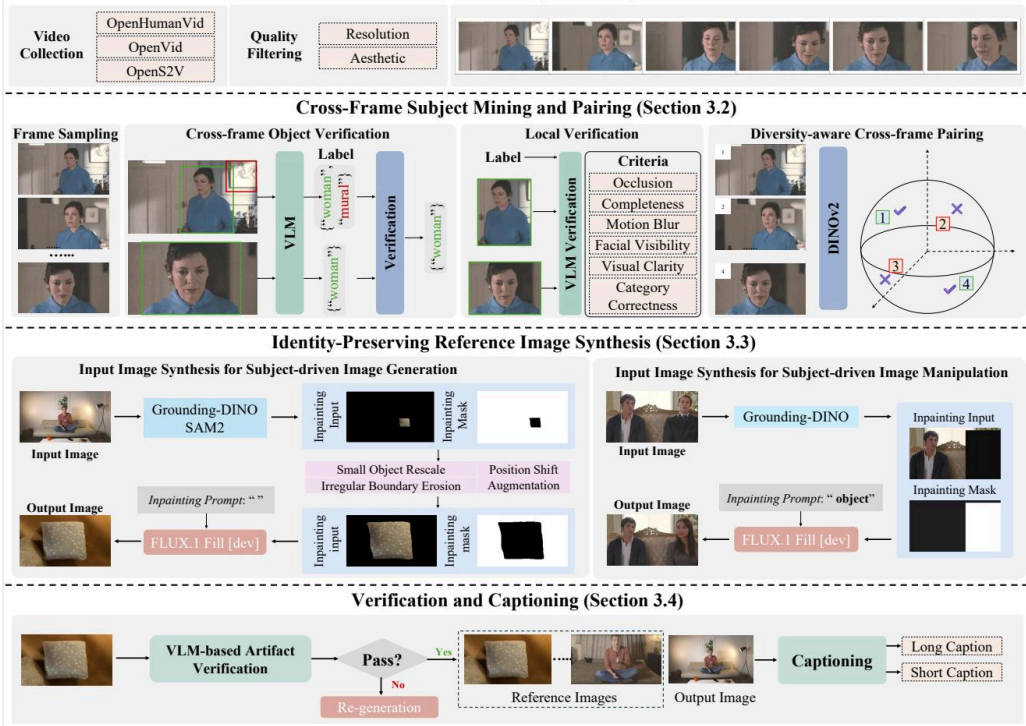
Use a VLM to enforce subject category consensus across frames, localize with Grounding-DINO, and select diverse pairs based on DINOv2 embedding distance.

03 **3. Identity-Preserving Synthesis**

Synthesize inputs for generation via mask-guided outpainting and for manipulation via box-guided inpainting using FLUX.1 Fill.

04 **4. Verification & Captioning**

Employ a VLM to check for artifacts, re-synthesize failed samples, and generate short and long captions for training.



Pipeline Deep Dive: Synthesis and Statistics

Subject-Driven Generation

- Inputs and targets are from different frames of the same video clip.
- Mask-guided outpainting completes the context around the subject to create the input image.

Subject-Driven Manipulation

- Edits are performed within a single frame.
- Box-guided inpainting erases a target object, creating an input for replacement while preserving the background.

Dataset Statistics

- 2.5M total samples and 4.35M images.
- Spans four sub-tasks: single/multi-subject generation and single/multi-subject manipulation.
- Broad coverage of categories including people, objects, and diverse environments.

OSBench: A New Evaluation Benchmark

Motivation

- Prior benchmarks often focus on clean, single-subject portraits and neglect complex scenes or manipulation tasks.

Four Sub-tasks

- Single-subject Generation
- Multi-subject Generation
- Single-subject Manipulation
- Multi-subject Manipulation

VLM-based Evaluation

- Uses GPT-4.1 as a judge with rubricized prompts to score outputs on multiple dimensions.
- **Generation:** Prompt Adherence (PA), Identity Fidelity (IF).
- **Manipulation:** Manipulation Fidelity (MF), Background Consistency (BC).

OSBench: Quantitative Results

- Most open-source models exhibit very low performance on manipulation tasks, indicating weak identity replacement and background preservation.
- Even strong closed-source models show a marked performance drop from generation to manipulation, especially in multi-subject cases.

Method	Size (B)	Gen Single PA	Gen Single IF	Gen Single Overall	Gen Multi PA	Gen Multi IF	Gen Multi Overall	Manip Single MF	Manip Single BC	Manip Single Overall	Manip Multi MF	Manip Multi BC
Gemini 2.5 Flash Image Preview	-	9.07	8.83	8.92	8.17	8.42	8.14	7.23	8.20	7.16	6.90	4.87
GPT-4o-2024-11-20	-	9.23	8.49	8.82	8.39	7.59	7.82	7.64	6.22	6.80	7.10	2.93
Gemini 2.0 Flash Image Preview	-	8.33	8.30	8.27	7.62	7.38	7.39	3.37	3.22	2.25	4.98	2.38
UNO	12	7.53	5.30	5.95	7.13	6.12	6.40	2.83	0.55	0.78	1.45	0.85
DreamO	12	7.10	4.12	5.13	7.03	6.18	6.52	3.86	0.27	0.38	2.13	0.71
XVerse	12	6.58	1.95	3.11	7.00	5.28	5.95	3.15	0.38	0.65	0.82	0.67
DreamOmni2	19	8.00	7.35	7.53	7.42	6.62	6.88	6.08	6.30	5.72	5.10	1.40
Qwen-Image-Edit-2509	20	9.03	8.55	8.77	7.83	7.25	7.44	7.63	6.53	6.69	7.90	3.85
OmniGen2	7	8.50	8.70	8.55	7.98	7.97	7.89	5.92	5.00	4.99	7.37	3.27

Ablation: Impact of Fine-tuning with OpenSubject

- Fine-tuning on OpenSubject consistently improves OmniGen2's performance, raising the average score from 6.43 to 7.22.
- The most significant gains are in manipulation (Single-subject +0.81, Multi-subject +1.91), showing the value of our specialized training data.
- Identity fidelity in generation is also enhanced without sacrificing prompt adherence.

Method	Single-Gen PA	Single-Gen IF	Single-Gen Overall	Multi-Gen PA	Multi-Gen IF	Multi-Gen Overall	Single-Manip MF	Single-Manip BC	Single-Manip Overall	Multi-Manip MF	Multi-Manip BC	M M O
OmniGen2 (baseline)	8.50	8.70	8.55	7.98	7.97	7.89	5.92	5.00	4.99	7.37	3.27	4
+ T2I	8.62 (+0.12)	8.17 (-0.53)	8.33 (-0.22)	8.12 (+0.14)	7.72 (-0.25)	7.81 (-0.08)	5.08 (-0.84)	3.22 (-1.78)	3.60 (-1.39)	6.47 (-0.90)	1.97 (-1.30)	2 (
+ OpenSubject (ours)	8.30 (-0.20)	8.95 (+0.25)	8.58 (+0.03)	8.05 (+0.07)	8.65 (+0.68)	8.26 (+0.37)	7.18 (+1.26)	5.22 (+0.22)	5.80 (+0.81)	7.85 (+0.48)	5.20 (+1.93)	6 (

Evaluation on External Benchmark: OmniContext

- Fine-tuning on OpenSubject improves performance on the OmniContext benchmark, especially in settings requiring integration of multiple subjects or strong scene grounding.

Model	Size (B)	SINGLE Character	SINGLE Object	MULTIPLE Character	MULTIPLE Object	MULTIPLE Char. + Obj.	SCENE Character	SCENE Object	SCENE Char. + Obj.	Average↑
FLUX.1 Kontext [Max]	-	8.48	8.68	-	-	-	-	-	-	-
Gemini 2.5 Flash Image Preview	-	8.52	9.14	7.80	8.64	6.63	6.74	7.11	6.04	7.58
GPT-4o	-	8.90	9.01	9.07	8.95	8.54	8.90	8.44	8.60	8.80
Baseline (OmniGen2)	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)

Evaluation on External Benchmark: ImgEdit

- Training with OpenSubject also boosts performance on the ImgEdit benchmark for instruction-based editing.
- The largest gains are in categories requiring precise localization like 'Extract', 'Hybrid', and 'Add', demonstrating improved edit robustness.

Model	Add	Adjust	Extract	Replace	Remove	Background	Style	Hybrid	Action	Overall ↑
FLUX.1 Kontext [Pro]	4.25	4.15	2.35	4.56	3.57	4.26	4.57	3.68	4.63	4.00
GPT-Image-1 [High]	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	4.65	4.34	3.69	4.49	4.65	4.32	4.13	3.66	4.59	4.28
Baseline (OmniGen2)	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)	3.72 (+0.28)

Qualitative Comparison

- Models often struggle to preserve identity in multi-subject scenes and can modify regions outside the intended edit area.
- Our model, fine-tuned on OpenSubject, demonstrates superior performance in these challenging scenarios.

Key Contributions

- **OpenSubject Corpus:** A large-scale, 2.5M-sample dataset derived from videos, designed for subject-driven generation and manipulation.
- **Scalable Pipeline:** A four-stage, automated pipeline for data construction that ensures subject consistency, diversity, and scalability.
- **Open-ended Benchmark:** OSBench, a comprehensive benchmark with four sub-tasks and a rubricized VLM judge for nuanced evaluation.
- **Improved Models:** Demonstrated that training on OpenSubject significantly enhances identity fidelity and manipulation robustness, particularly in multi-subject settings.

Ethical Considerations

- All data is sourced from publicly available, open-licensed corpora, in compliance with original licensing terms.
- The dataset will be released for research-only purposes.
- An acceptable-use policy will prohibit applications such as biometric identification, surveillance, and non-consensual impersonation.

Thank You

- Project Page: https://github.com/LAW1223/OpenSubject