OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation

✨ NeurIPS D&B 2025 ✨

Shenghai Yuan[1,4], Xianyi He[1,4], Yufan Deng[1], Yang Ye[1,4], Jinfa Huang[3], Bin Lin[1,4], Jiebo Luo[3], Li Yuan[1,2,†]

[1]Peking University, [2]Peng Cheng Laboratory, [3]University of Rochester, [4]Rabbitpre Intelligence

Introduction

北京大学 PEKING UNIVERSITY

Development of foundational model

sora CogVideo OpenSora Plan 腾讯混元-Video

Downstream fine-tuning

Trajectory

Novel View

Pose

Depth

TORA ViewCrafter AnimateAnyone DepthCrafter

Introduction

What is subject-to-video generation

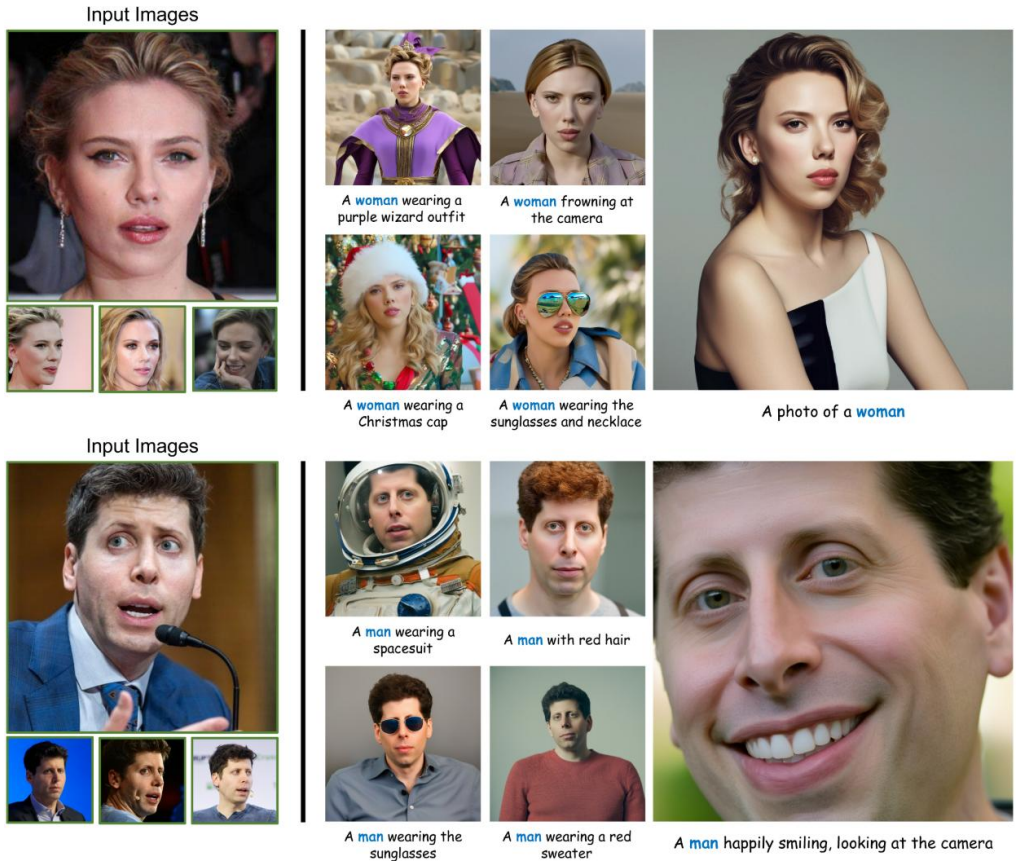
- Customized requirements
- Artistic Creation

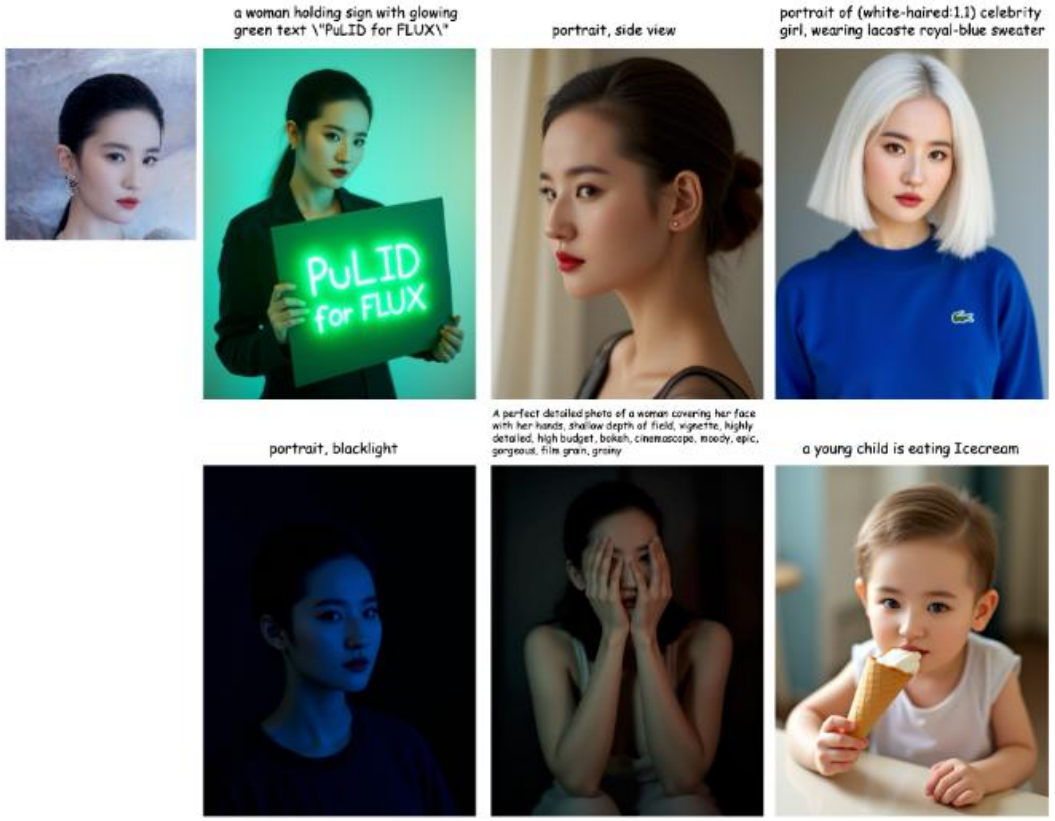
Introduction

image/video subject-to-video generation

Image:
PhotoMaker,
InstantID,
PuLID,
IP-Adapter,
…

Video: *???*





Motivation

Key Challenges for Subject-to-Video

Challenges 1

Poor generalization: These models often perform poorly when encountering subject categories not seen during training. For instance, a model trained exclusively on Western subjects typically performs worse when generating Asian subjects.

Challenges 2

Copy-paste issue: The model tends to directly transfer the pose, lighting, and contours from the reference image to the video, resulting in unnatural outcomes.

Challenges 3

Inadequate human fidelity: Current models often struggle to preserve human identity as effectively as they do non-human entities.

OpenS2V-Eval

Inadequacy of existing Benchmarks

Table 1: **Comparison of the Characteristics of our OpenS2V-Eval with existing Benchmarks** Most of them focus on T2V and neglect the evaluation of subject naturalness. _ means suboptimal.

Benchmark	# Type	Visual Quality	Text Relevance	Motion	Subject Consistency	Subject Naturalness
Make-a-Video-Eval [81]	Text-to-Video	✓	✓	✗	✗	✗
FETV [58]	Text-to-Video	✓	✓	✓	✗	✗
T2VScore [100]	Text-to-Video	✓	✓	✓	✗	✗
EvalCrafter [57]	Text-to-Video	✓	✓	✓	✗	✗
VBench [36]	Text-to-Video	✓	✓	✓	✗	✗
VBench++ [37]	Text-to-Video	✓	✓	✓	✗	✗
ChronoMagic-Bench [115]	Text-to-Video	✓	✓	✓	✗	✗
ConsisID-Bench [113]	Subject-to-Video	✓	✓	✓	✓	✗
Alchemist-Bench [13]	Subject-to-Video	✓	✓	✓	✓	✗
A2 Bench [21]	Subject-to-Video	✓	✓	✓	✓	✗
VACE-Bench [40]	Subject-to-Video	✓	✓	✓	✓	✗
OpenS2V-Eval	Subject-to-Video	✓	✓	✓	✓	✓

OpenS2V-Eval

Construction pipeline

Figure 2: **The Pipeline of Constructing OpenS2V-Eval.** (Left) Our benchmark includes not only real subject images but also synthetic images constructed through GPT-Image-1 [1], allowing for a more comprehensive evaluation. (Right) The metrics are tailored for subject-to-video generation, evaluating not only S2V characteristics (e.g., consistency) but also basic video elements (e.g., motion).

OpenS2V-5M

Inadequacy of existing Datasets

Table 2: **Comparison of the Statistics of OpenS2V-5M with existing Video Generation Datasets** Most of them are inadequate for extending foundational models to subject-to-video generation task

Dataset	# Type	Resolution	Video Clips	Average Length (s)	Video Duration (h)
MSRVTT [106]	Text-to-Video	240P	10K	14.4	40
WebVid-10M [4]	Text-to-Video	360P	10M	18.7	52K
InternVid [95]	Text-to-Video	720p	234M	11.7	760K
HD-VG-130M [94]	Text-to-Video	720p	130M	4.9	178K
Panda-70M [12]	Text-to-Video	720P	70M	8.6	167K
OpenVid-1M [67]	Text-to-Video	512P	1M	7.2	2K
Koala-36M [91]	Text-to-Video	720P	36M	17.2	172K
ChronoMagic-Pro [115]	Text-to-Video	720p	460K	234.8	30K
OpenHumanVid [45]	Text-to-Video	720P	52.3M	4.9	70K
OpenS2V-5M	Subject-to-Video	720P	5.4M	6.6	10K

OpenS2V-5M

Construction pipeline

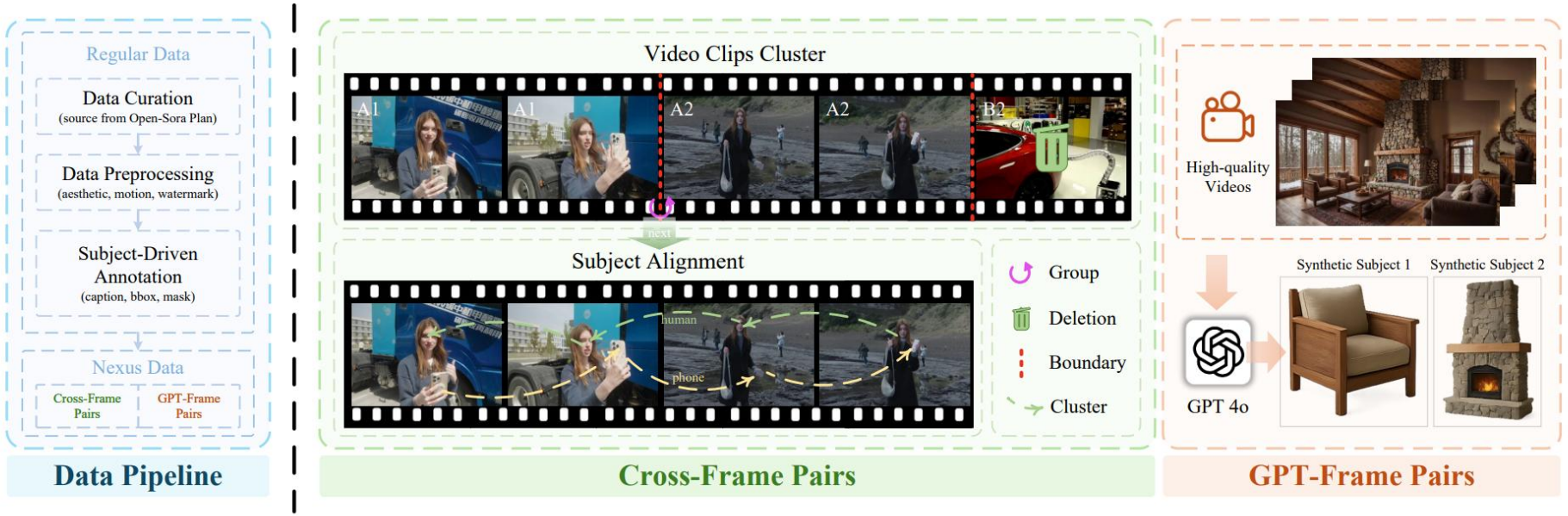
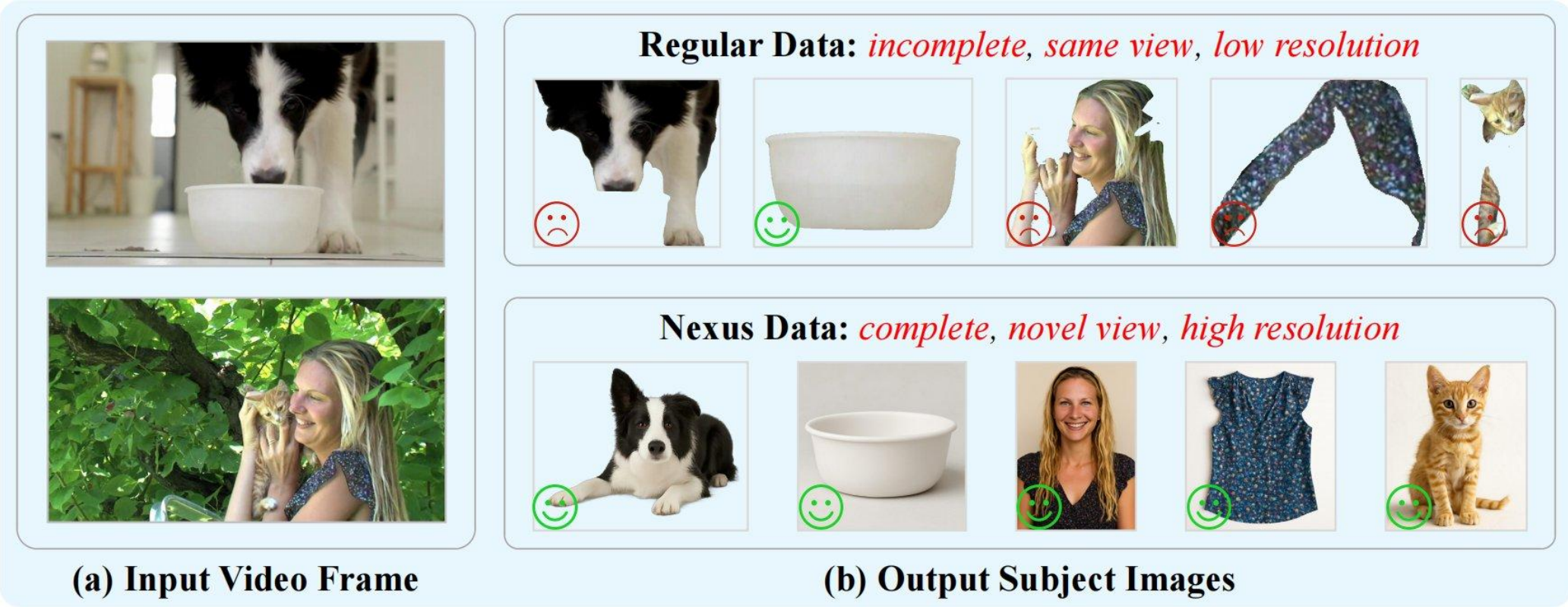


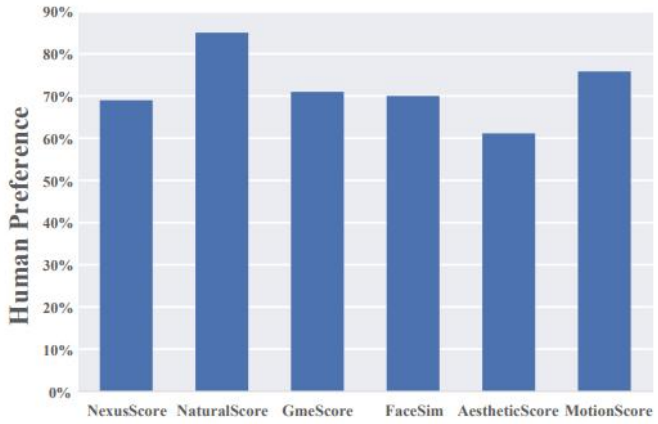
Figure 4: **The Pipeline of Constructing OpenS2V-5M.** First, we filter low-quality videos based on scores such as aesthetics and motion, then utilize GroundingDino [56] and SAM2.1 [76] to extract subject images and get Regular Data. Subsequently, we create Nexus Data through cross-video association and GPT-Image-1 [1] to address the three core issues encountered by S2V models.

OpenS2V-5M

Inadequacy of existing Datasets



Experiment



(a) Verification of Automatic Metric

(b) Verification of the Proposed Dataset

Figure 9: **(a) Alignment between Automatic Metrics and Human Perception.** The proposed metrics are comparable to other metrics [17, 6, 16] in terms of human preference. **(2) Validation of ConsisID-Nexu-5M with † and without ‡ Nexus Data.** Training are based on ConsisID [113].

Experiment

Figure 6: **Qualitative Comparison among Different Methods for the Open-Domain Subject-to-Video task.** Existing methods handle non-human entities better than human identities, and perform better with single subject compared to multiple subjects.

Experiment

Figure 7: **Qualitative Comparison among Different Methods for the Human-Domain Subject-to-Video task.** They are unable to generate consistent side profiles and suffer from copy-paste issues.

Experiment

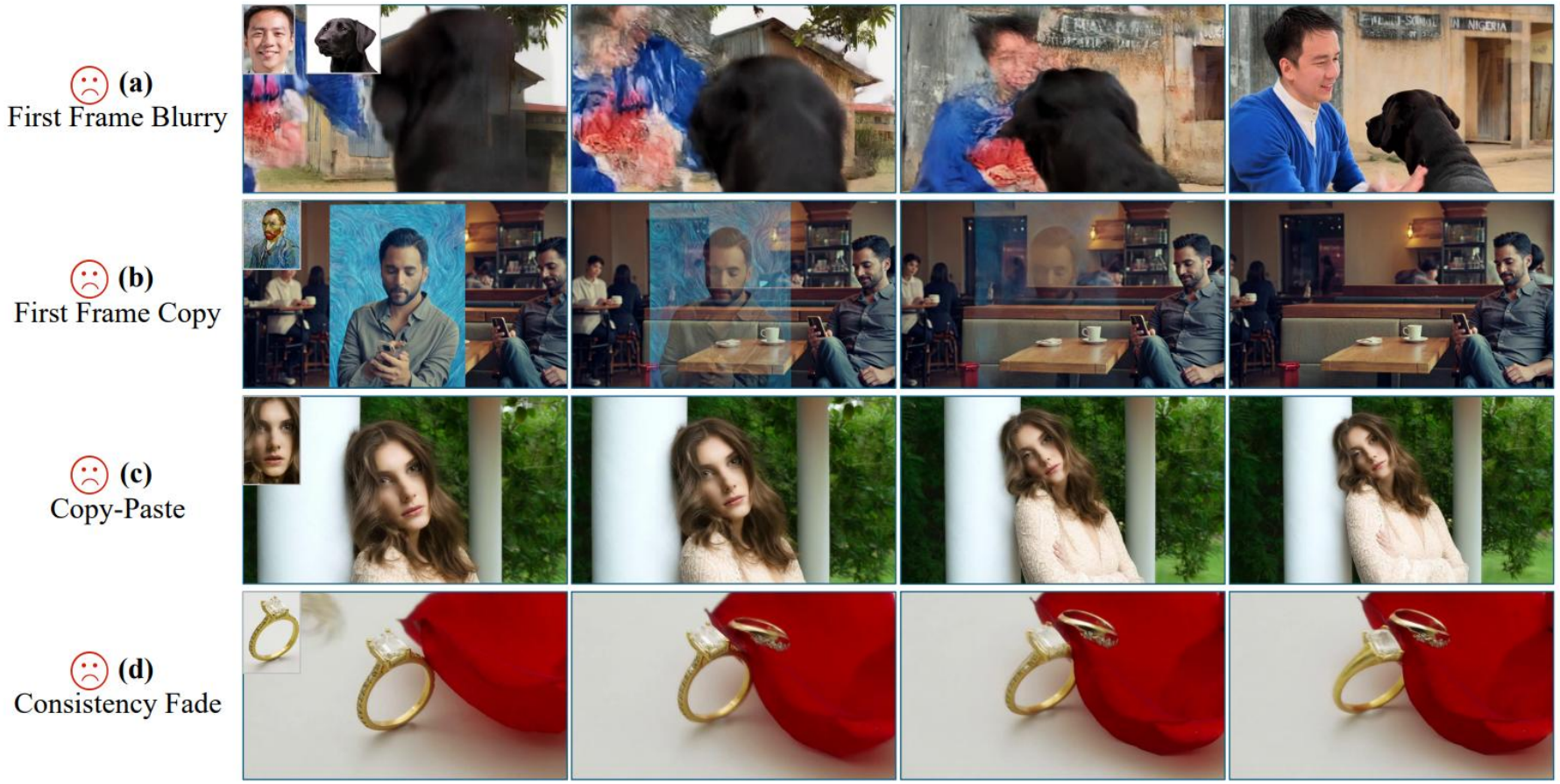


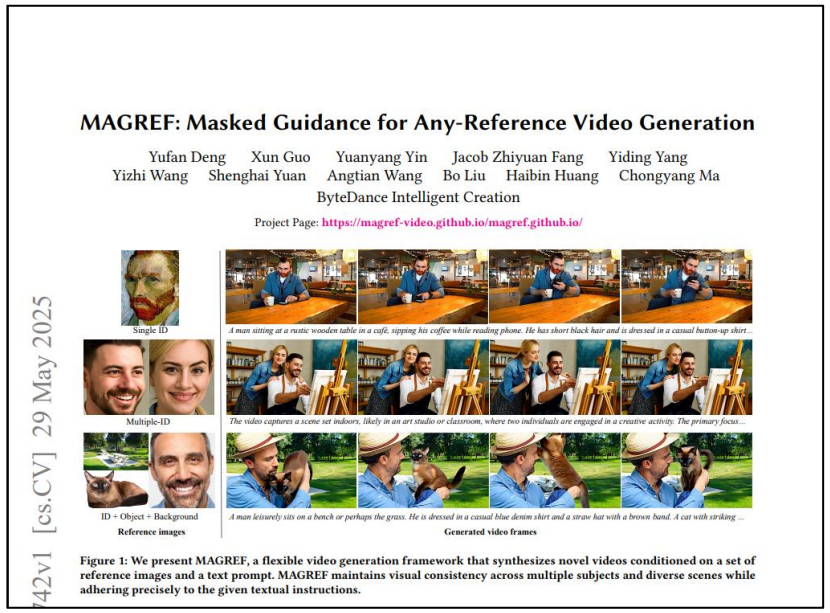
Figure 13: **Example of Common Issues faced by current Subject-to-Video Generation Models.** These videos are generated by Kling [43] and SkyReels-A2 [21] for demonstration purposes only.

Follow Up

BestWishYsh/OpenS2V-5M
Updated 12 days ago • 40.5k • 12

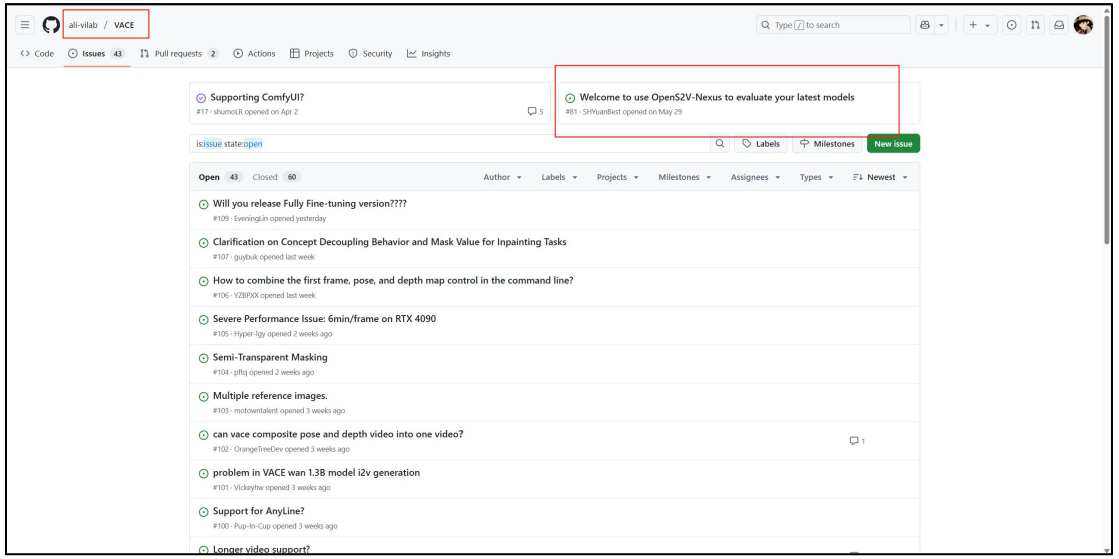
Reach ~40,000 downloads on Hugging Face in just one month

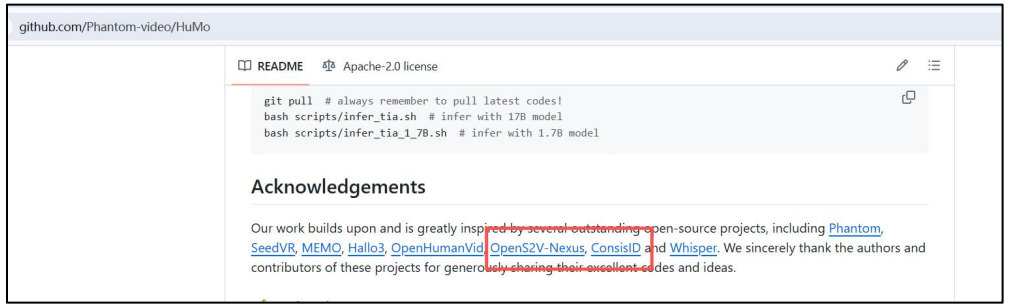
[1] Vace: All-in-one video creation and editing[J].



[2] MAGREF: Masked Guidance for Any-Reference Video Generation.

[3] HuMo: Human-Centric Video Generation via Collaborative Multi-Modal Conditioning.





Conclusion

- **OpenS2V-Eval**: The most comprehensive S2V evaluation benchmark in the field, featuring 180 multi-domain prompts paired with dual-category test data (real/synthetic). We propose NexusScore (subject consistency), NaturalScore (naturalness), and GmeScore (text-video alignment) for precise multidimensional capability assessment.

- **OpenS2V-5M Dataset**: A newly open-sourced collection of 5.4 million 720P HD <image-text-video> triplets. Through cross-video relational segmentation and multi-perspective synthesis techniques, it achieves exceptional thematic diversity and annotation quality.

- **Novel Insights for S2V Model Selection**: Our evaluation framework enables comprehensive benchmarking of 18 leading S2V models, revealing comparative advantages across complex scenarios.

ConsisID

OpenS2V-Nexus

Personal Page

Thank you!

[4] https://github.com/PKU-YuanGroup/OpenS2V-Nexus