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

Shenghai Yuan, Xinyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, Li Yuan

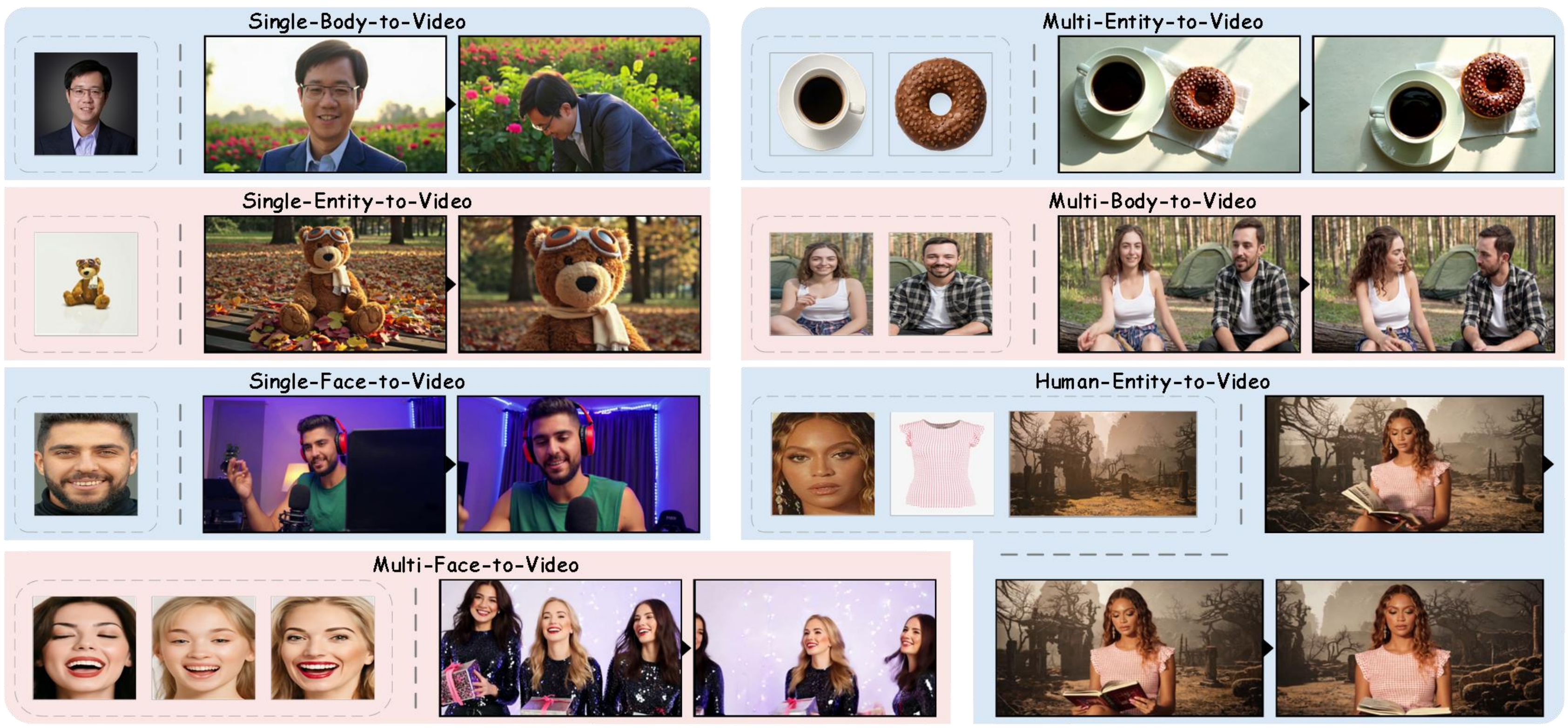
Page Code

NEURAL INFORMATION PROCESSING SYSTEMS

Subject-to-Video Generation

Highlight:

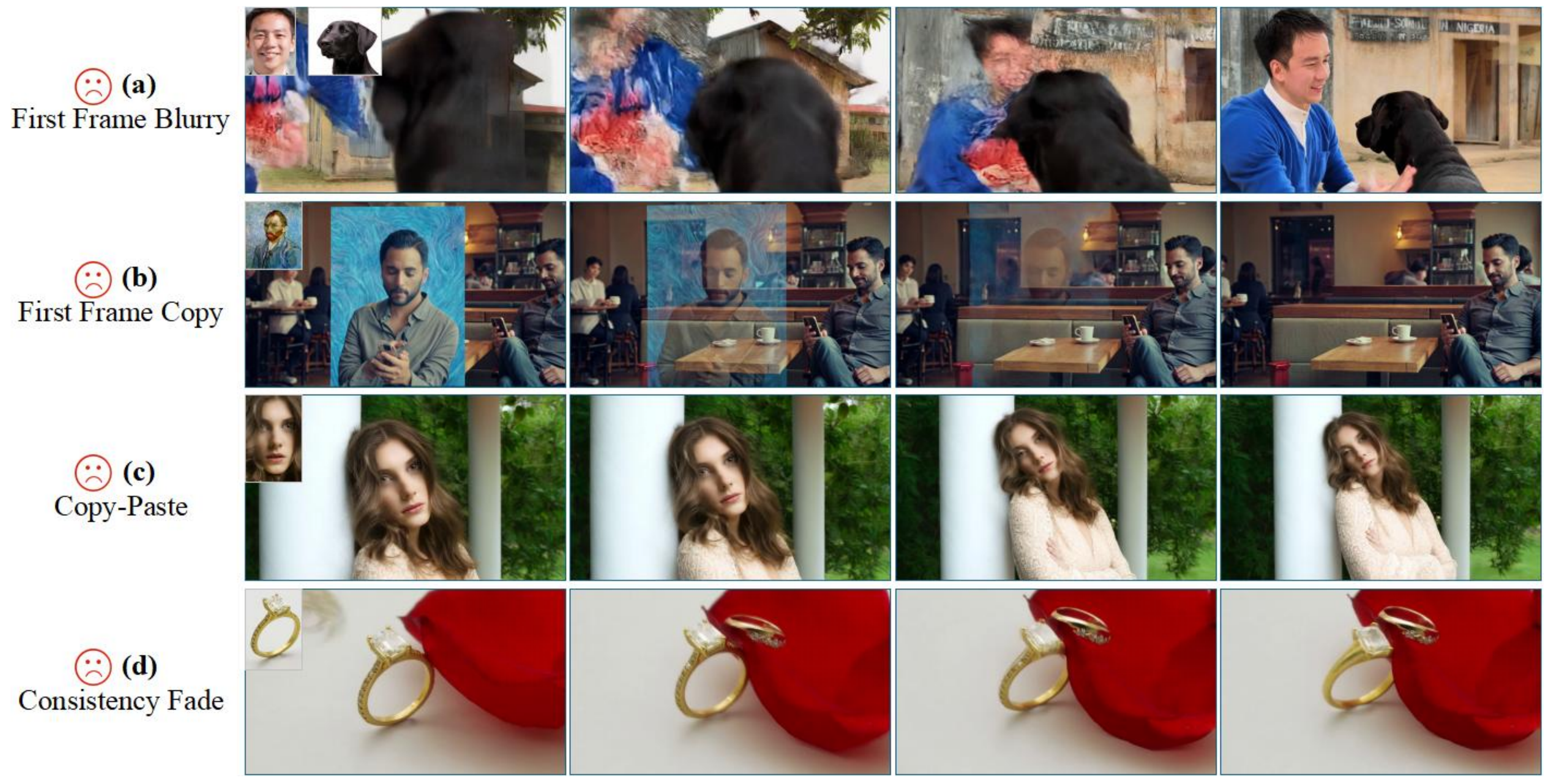
- **The First Comprehensive S2V Benchmark. (fully open-sourced)**
- **The First Million-Scale S2V Datasets. (fully open-sourced)**



Subject-to-Video (S2V) aims to create videos that faithfully incorporate reference content, providing enhanced flexibility in the production of videos.

Key Challengs for S2V Models

- **Poor generalization:** These models often perform poorly when encountering subject categories not seen during training.
- **Copy-paste issue:** The model tends to directly transfer the pose, lighting, and contours from the reference image to the video.
- **Inadequate human fidelity:** Current models often struggle to preserve human identity as effectively as non-human entities.

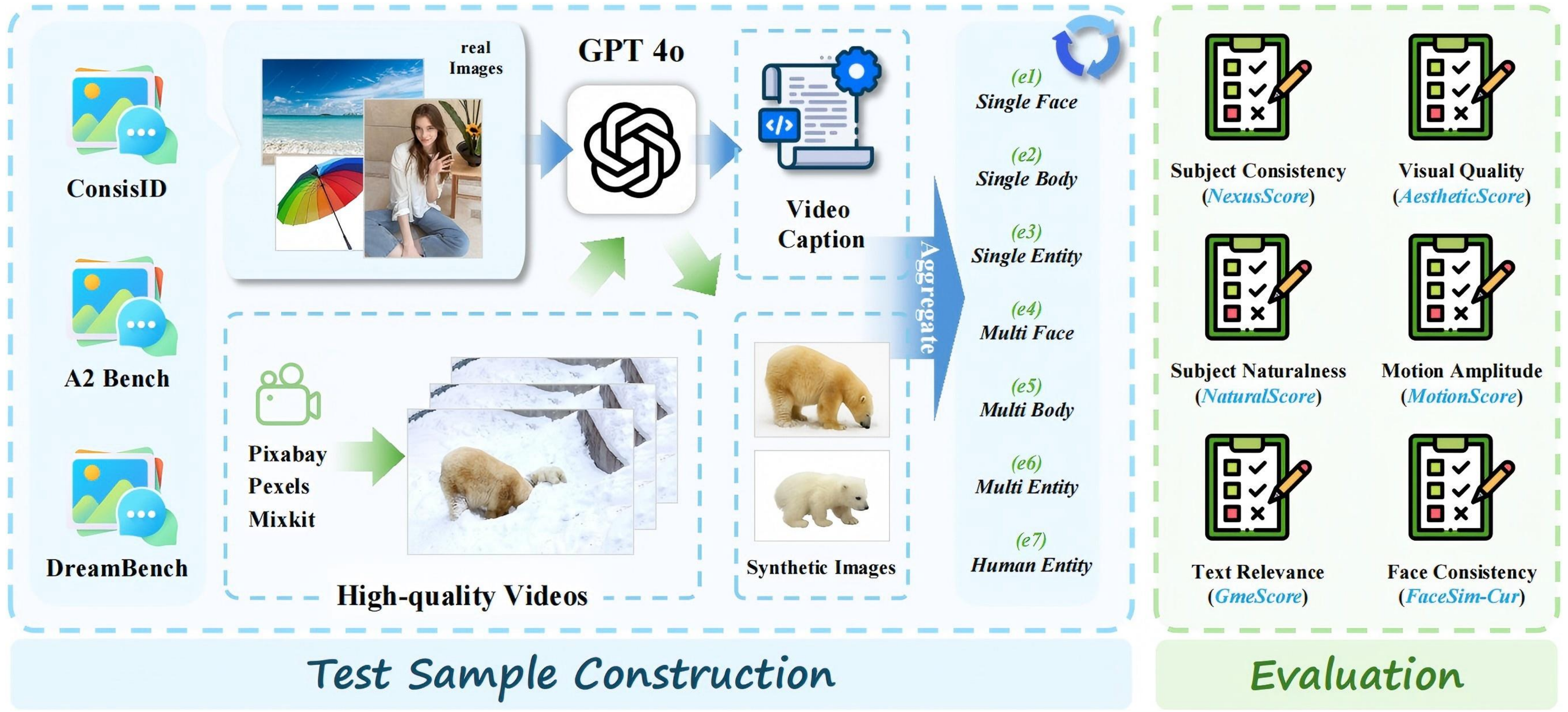


In addition to the three key challenges outlined before, we also observe some noteworth phenomenon, as shown above (e.g., First Frame Blurry).

Contact: yuanshenghai@stu.pku.edu.cn

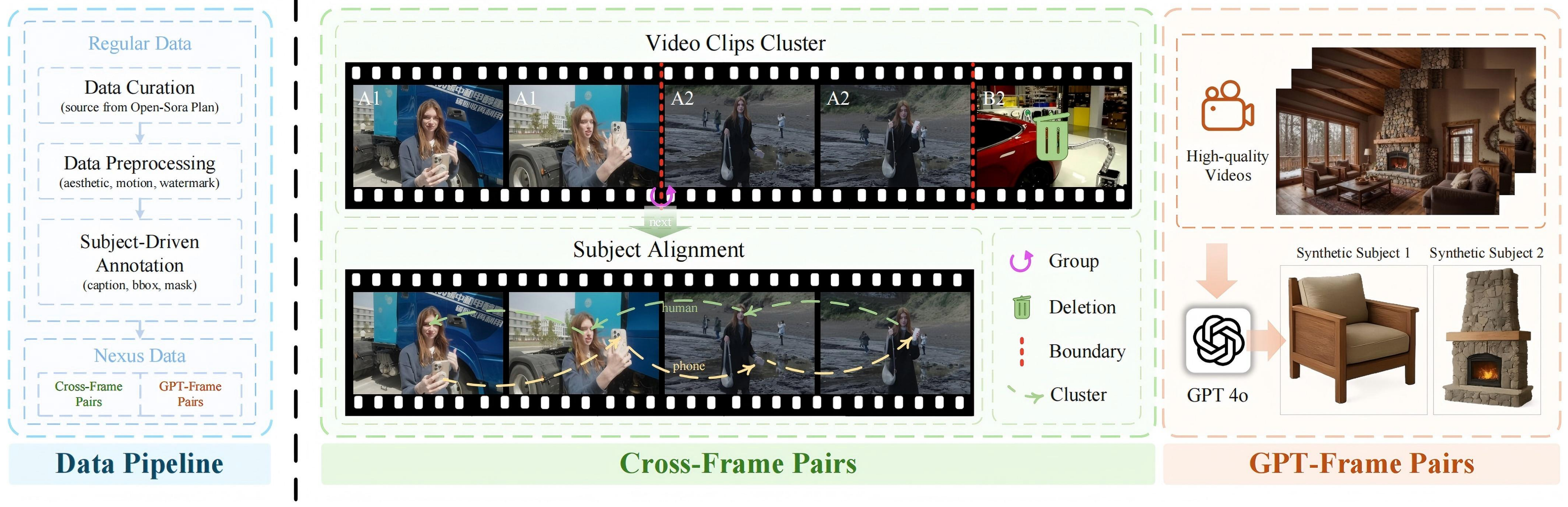
OpenS2V-Eval Pipeline

Our benchmark includes not only **real** subject images but also **synthetic** images



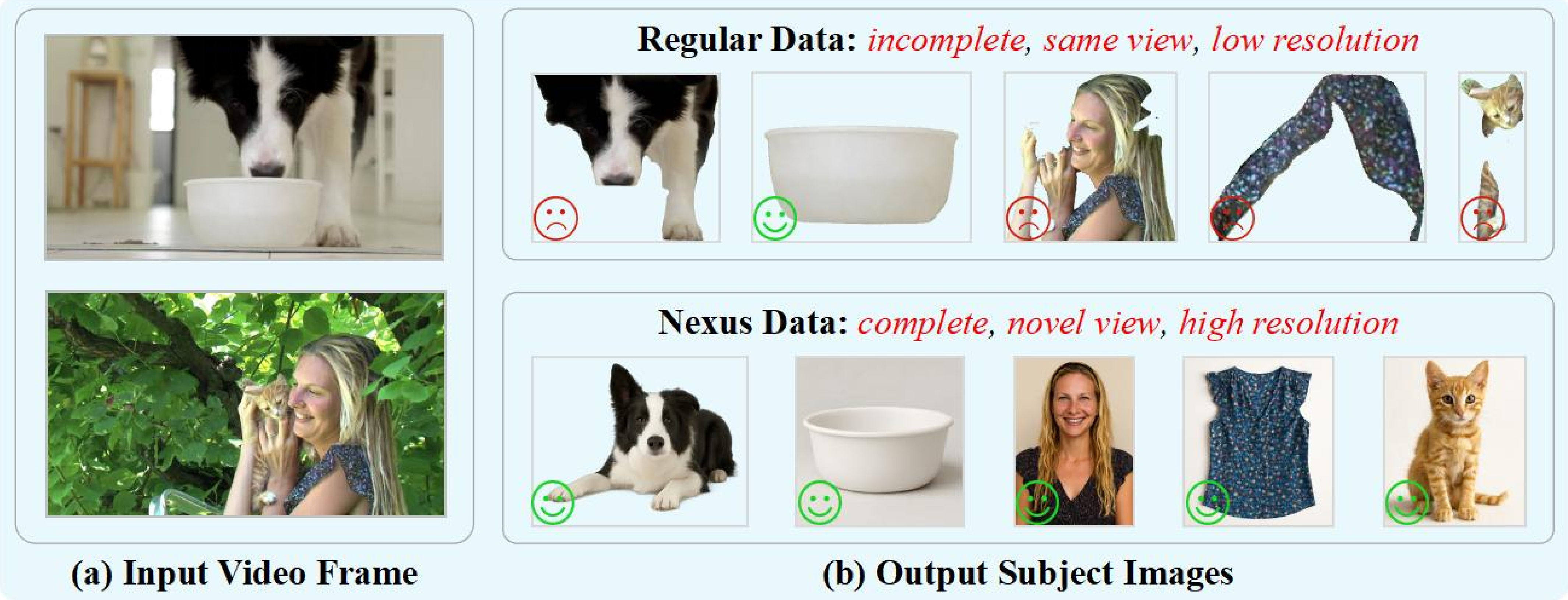
OpenS2V-5M Pipeline

We create data through **cross-video association** and **GPT-Image-1** to address the three core issues



Regular Data *vs* OpenS2V-5M

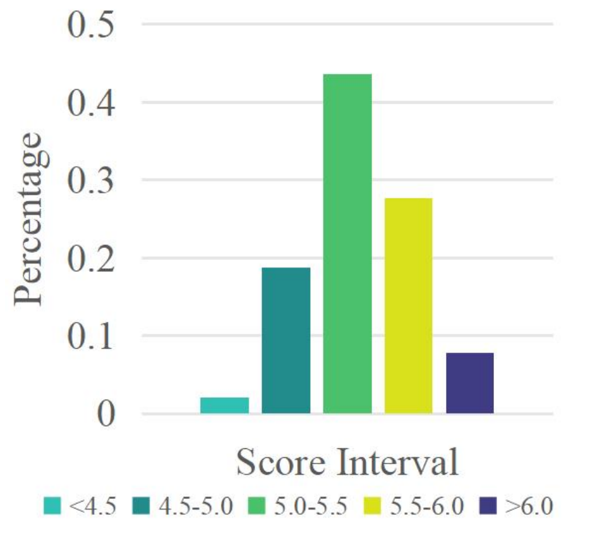
Compared to Regular Data, our **Nexus Data** is of higher quality.


Regular Data: incomplete, same view, low resolution
Nexus Data: complete, novel view, high resolution

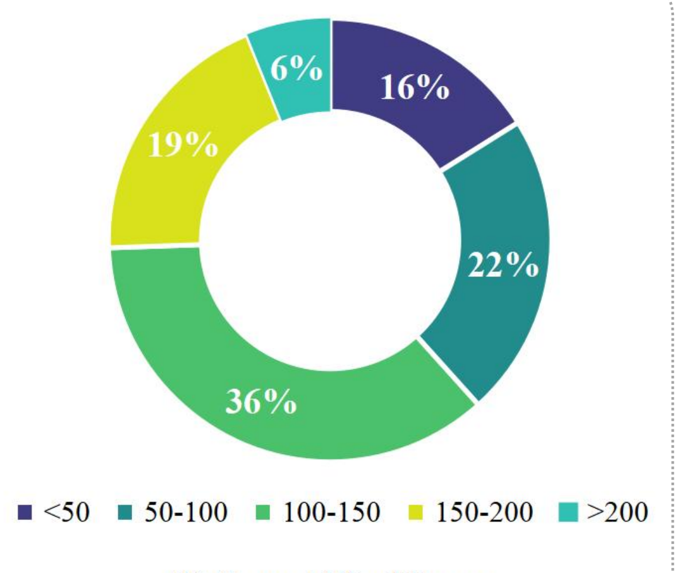
(a) Input Video Frame **(b) Output Subject Images**

Statistic

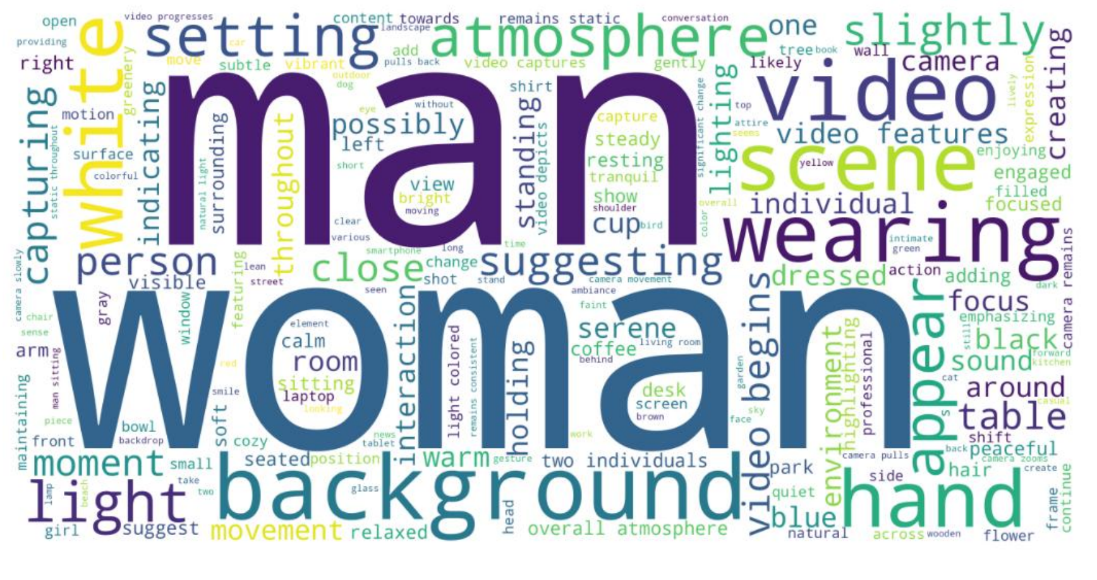
OpenS2V covers diverse categories and prompt words, with subject images displaying high aesthetics



(a) Distribution of Aesthetic Scores



(b) Prompt Word Range



(c) Prompt Word Cloud

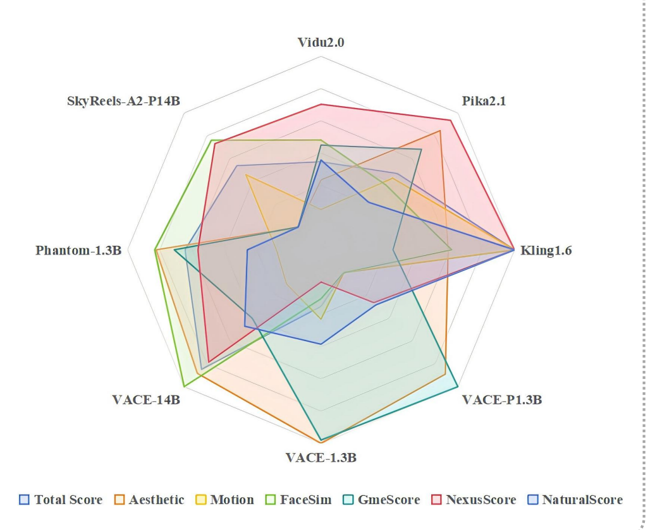
(d) Subject Categories

Results

For simplicity, only total score is shown here.

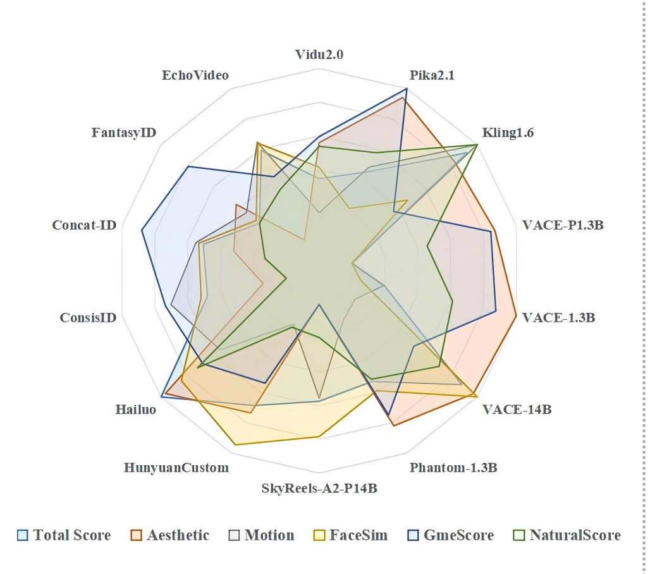
Method	Venue	Total Score↑
Vidu2.0 [5]	Closed-Source	51.95%
Pika2.1 [46]	Closed-Source	51.88%
Kling1.6 [45]	Closed-Source	56.23%
VACE-P1.3B [42]	Open-Source	48.98%
VACE-1.3B [42]	Open-Source	49.89%
VACE-14B [42]	Open-Source	**57.55%**
Phantom-1.3B [58]	Open-Source	54.89%
Phantom-14B [58]	Open-Source	56.77%
SkyReels-A2-P14B [22]	Open-Source	52.25%
MAGREF-480P [19]	Open-Source	52.51%

Open-Domain



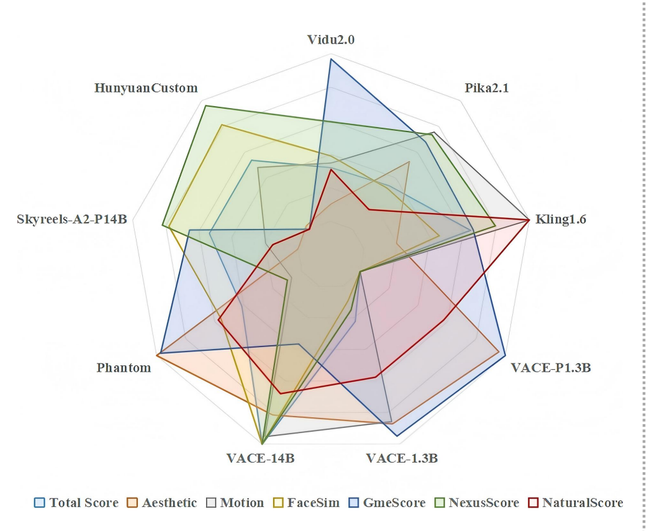
Method	Venue	Domain	Total Score↑
Vidu2.0 [5]	Closed-Source	Open-Domain	57.70%
Pika2.1 [46]	Closed-Source	Open-Domain	56.84%
Kling1.6 [45]	Closed-Source	Open-Domain	60.19%
VACE-P1.3B [42]	Open-Source	Open-Domain	53.97%
VACE-1.3B [42]	Open-Source	Open-Domain	54.90%
VACE-14B [42]	Open-Source	Open-Domain	**65.78%**
Phantom-1.3B [58]	Open-Source	Open-Domain	60.00%
Phantom-14B [58]	Open-Source	Open-Domain	64.22%
SkyReels-A2-P14B [22]	Open-Source	Open-Domain	56.43%
HunyuanCustom [35]	Open-Source	Open-Domain	61.22%
MAGREF-480P [19]	Open-Source	Open-Domain	57.72%
Hailuo [90]	Closed-Source	Human-Domain	65.26%
ConsisID [119]	Open-Source	Human-Domain	54.19%
Concat-ID-CogVideoX [129]	Open-Source	Human-Domain	55.89%
Concat-ID-Wan-AdaLN [129]	Open-Source	Human-Domain	59.85%
FantasyID [126]	Open-Source	Human-Domain	54.33%
EchoVideo [100]	Open-Source	Human-Domain	56.36%
VideoMaker [107]	Open-Source	Human-Domain	54.23%
ID-Animator [31]	Open-Source	Human-Domain	49.75%
Ours †	-	Human-Domain	58.00%
Ours ‡	-	Human-Domain	59.23% (+1.23%)

Human-Domain



Method	Venue	Total Score↑
Vidu2.0 [5]	Closed-Source	52.90%
Pika2.1 [46]	Closed-Source	53.12%
Kling1.6 [45]	Closed-Source	56.67%
VACE-P1.3B [42]	Open-Source	49.20%
VACE-1.3B [42]	Open-Source	51.13%
VACE-14B [42]	Open-Source	**61.75%**
Phantom-1.3B [58]	Open-Source	54.50%
Phantom-14B [58]	Open-Source	57.02%
SkyReels-A2-P14B [22]	Open-Source	55.06%
HunyuanCustom [35]	Open-Source	56.89%
MAGREF-480P [19]	Open-Source	53.44%

Single-Domain


Vidu2.0
Pika2.1
Kling1.6
VACE-P1.3B
VACE-1.3B
Phantom
HunyuanCustom
Skyreels-A2-P14B
Total Score
Aesthetic
Motion
FaceSim
GmeScore
NexusScore
NaturalScore