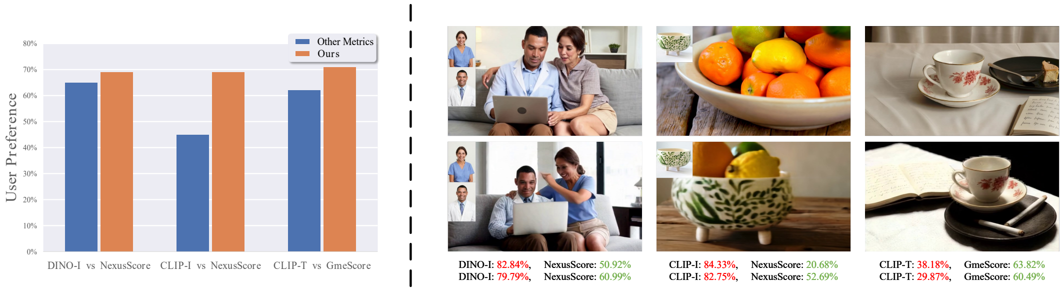
Paper Appendix for *OpenS2V-Nexus*: A Ultra-Scale Dataset and Benchmark for Subject-Consistent Video Generation

A Related Works: Subject-Consistency Video Generation Models

Diffusion models are widely acknowledged for their remarkable generative capabilities [77, 76, 74, 65, 66, 64, 86, 114, 24], which have significantly advanced the development of subject-consistency generation models [39, 28, 27, 10]. Initially, researchers utilized tuning-based methods to generate consistent image content, such as DreamBooth [79], Lora [31], and Textual Inversion [25]. These methods integrate specific reference content into the training process through fine-tuning existing parameters, adding extra parameters, or modifying text embeddings. Later models, including Magic-Me [67], MotionBooth [104], and DreamVideo [100], extended these approaches to video generation. However, since these methods require training on each new reference content before inference, their practical application is limited. To mitigate the high computational cost, tuning-free methods were



(a) Quantitative Comparison with existing Metrics

(b) Qualitative Comparison with existing Metrics

Figure 10: **Comparison with Existing Metircs for Subject Consistency and Text Relevance.** The proposed automatic metricsalign more closely with human preferences compared to the commonly used DINO-I [118], CLIP-I [75], and CLIP-T [75] in existing S2V methods [41, 57, 38, 22].

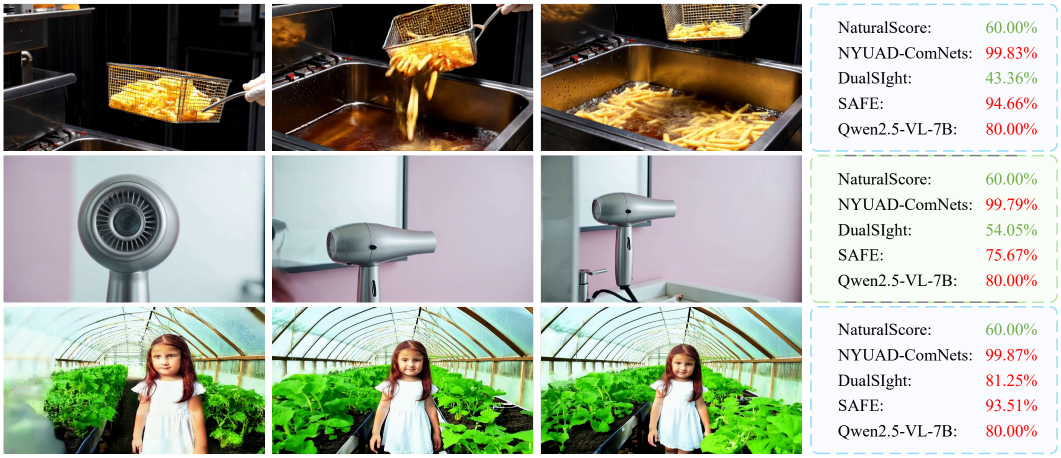


Figure 11: **Comparison with Existing Methods for Subject Naturalness.** Existing AIGC anomaly detection models and multimodal models are both prone to misidentifying generated content as real.

introduced. A notable example is IP-Adapter [112], which leverages large datasets to train additional adapters for open-domain subject-consistency generation. However, due to its lower fidelity to human identity, InstantID [94] and PhotoMaker [48] developed human-domain subject-consistency generation models based on this approach. Similar to these image consistency techniques, ID-Animator [30] and ConsisID [115] achieved tuning-free Subject-to-Video (S2V) generation on UNet and DiT, respectively. Nevertheless, these approaches [125, 99, 23, 122] are confined to the human domain, limiting their broader applicability. Recent works, such as Phantom [57], VACE [41], and SkyReels-A2 [22], have demonstrated the ability to generate consistent multi-subject videos in the open domain [50, 13, 36], gradually narrowing the gap with commercial S2V models [44, 45, 89, 5]. However, a unified and comprehensive benchmark to assess the strengths and weaknesses of these models remains absent, and the lack of publicly released training data impedes further progress in this field. Therefore, we introduce OpenS2V-Eval and OpenS2V-5M, aimed at bridging this gap.

B More Details of OpenS2V-Eval

B.1 Comparison with Existing Metircs for Subject Consistency and Text Relevance

As previously noted, Alchemist-Bench [13], VACE-Benchmark [41], and A2 Bench [22] enable the evaluation of open-domain S2V. However, these evaluations are typically derived from VBench [38] and are predominantly limited to global, coarse-grained assessments. Specifically, they often rely on CLIP [75] or DINO [118] to calculate the similarity between text and images, both of which have been shown to exhibit poor robustness [102, 117, 60]. To substantiate these claims, we employ an

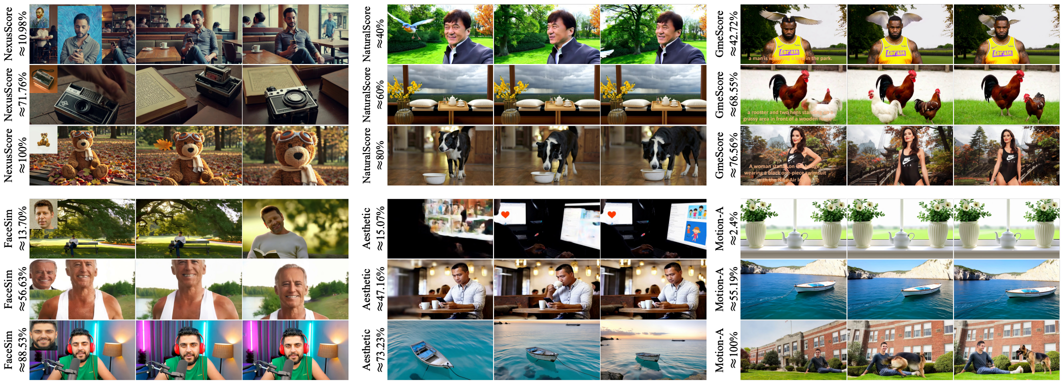


Figure 12: **Visual Reference for Varying Scores of Different Metircs.** It is evident that the proposed NexusScore, NaturalScore, and GmeScore are highly correlated with human perception.

evaluation akin to human evaluation to gather user preferences for DINO-I, CLIP-I, and CLIP-T. Additionally, six samples are randomly selected for qualitative analysis, as illustrated in Figure 10. The results demonstrate that the proposed NexusScore and GmeScore offer greater accuracy in assessing subject consistency and text relevance compared to others. All higher scores are better.

B.2 Comparison with Existing Metrics for Subject Naturalness

To evaluate whether a generated video is natural—meaning whether it complies with the laws of physics and common sense—a simple solution is to apply AIGC anomaly detection models [109, 47, 68, 2, 70], using the probability of the real label as the score. Alternatively, open-source multimodal large language models [3, 92, 52, 87] can be used for video scoring. However, we found that the former lacks accuracy, while the latter suffers from poor instruction-following performance and is prone to significant hallucinations. None of these methods perform as effectively as the NexusScore we propose, which is based on GPT-4o [1], as shown in Figure 11.

B.3 Visual Reference of Different Metrics

We also provide visual samples of NexusScore, NaturalScore, GmeScore, FaceSim-Cur [115], AestheticScore [16], and Motion-A [6] with different scoring scales, as shown in Figure 12. It can be observed that all the metrics are consistent with human perception, especially the three proposed automatic metrics targeting subject consistency, subject naturalness, and text relevance.

B.4 More Qualitative Analysis

We present further qualitative analysis, as illustrated in Figures 13, 22, 21, and 23. Both open-source and closed-source models encounter the following challenges:

Poor Generalization Although open-domain S2V models claim to support input from images of any category, they do not consistently produce satisfactory results. As illustrated in case 5 of Figure 21, while Kling [44] largely preserves the mole's body shape, it loses the original fur color. Other models [45, 57, 22] entirely lose the reference subject information. Furthermore, as the number of reference images increases, the model's ability to retain information progressively diminishes. This issue is particularly pronounced in open-source models [22, 41], as shown in cases 1–6 of Figure 21.

Copy-Paste Issue Existing models often inaccurately replicate the lighting, pose, expression, and other attributes from reference images directly onto generated videos, instead of generating content by learning the intrinsic features of the reference subjects. Although this may result in higher fidelity content, it generally fails to align with human perception and appears unnatural. As illustrated in Figure 13(c), the model directly places a face onto a person leaning against a pillar, creating an unnatural and visually awkward effect. This problem is particularly evident in generating human.

Inadequate Human Fidelity As demonstrated in Figures 21, 23, and 24, current models often face difficulties in preserving human identity as effectively as they preserve non-human entities.

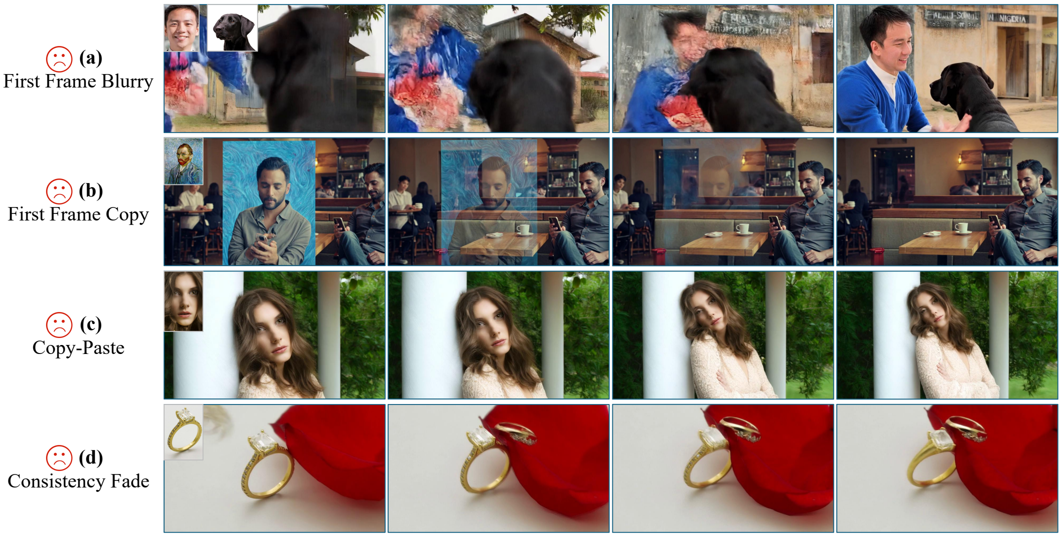
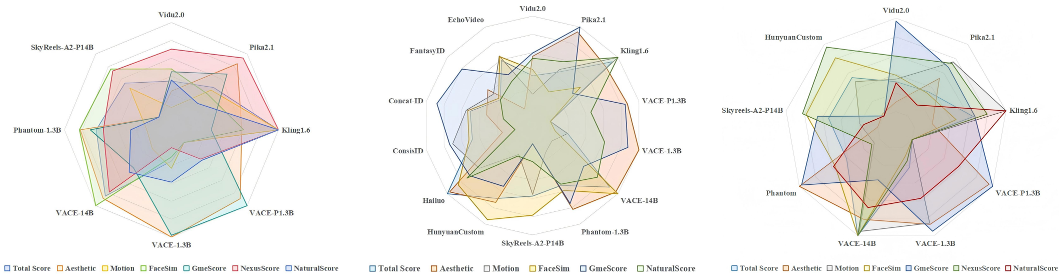


Figure 13: **Example of Common Issues faced by current Subject-to-Video Generation Models.** These videos are generated by Kling [44] and SkyReels-A2 [22] for demonstration purposes only.



(a) Open-Domain Evaluation

(b) Human-Domain Evaluation

(c) Single-Domain Evaluation

Figure 14: **Visualization of all the Quantitative Results in OpenS2V-Eval.**

While part of this issue can be attributed to human perception being more sensitive to facial changes, the primary cause lies in the models' insufficient capabilities. This is also one of the reasons why human-domain models exist, such as ConsisID [115], EchoVideo [99] and Hailuo [89].

First Frame Blurry or Copy In addition to the three core issues outlined above, we also observe a noteworthy phenomenon in which the model directly replicates the reference image into the generated video, as illustrated in Figure 13(b), generated by Kling [44]. Furthermore, it is possible that the first few frames of the generated video appear blurry, gradually becoming clear as shown in Figure 13(a), generated by SkyReels-A2 [22]. Similar phenomena are also observed in the Phantom [57], ConsisID [115], and Concat-ID [125] models, likely due to the use of VAE [11, 49] as the control signal.

Consistency Fade As shown in Figure 13(d), although the model effectively preserves both global and local information of the subject in the first half of the video, the diamond embedded in the ring gradually disappears as the sequence progresses. This issue may stem from the underlying video generation model [91, 42, 111], but it remains a noteworthy concern.

B.5 Guideline for Model Selection

We visualize all the results of OpenS2V-Eval, as shown in Figure 14. As the number of S2V models increases, the community faces challenges in selecting the most appropriate model, as each one tends to highlight its best results. To address this challenge, we offer model selection guidelines based on the evaluation outcomes of OpenS2V-Eval: (1) For content creators (e.g., advertisements, product displays), the closed-source Kling [44] is the clear leader, providing a more flexible and user-friendly experience. However, due to its high inference cost, more cost-effective alternatives such as Pika

[45] and Vidu [5] may be preferred. While these alternatives do not surpass Kling [44], they still outperform open-source models. (2) For community developers, it is recommended to base S2V model development on Phantom [57] or VACE [41], as it generates videos with relatively high quality and subject fidelity. Fine-tuning these methods can reduce development costs. (3) Although Hailuo has a narrower scope of application, it outperforms open-domain models like Kling in preserving human identity, making it more suitable for generating human-centric videos, such as those involving models and voice-over content. (4) For developing human-centric S2V models, open-source methods like HunyuanCustom [34], and ConsisID [125] offer high-quality pretrained weights, which may could also be extended to open-domain subject-to-video generation.

C More Details of OpenS2V-5M

C.1 Additional Details of Subject-Driven Processing

Human-Centric Filtering. Our data comes from 14,818,489 raw videos crawled from Internet through the Open-Sora Plan [51], consisting of no transition, clean clips with detailed raw captions. We design 100 human-related verbs and nouns as search terms, which lead to the identification of 12,654,783 human-related videos based on the raw captions. Finally, we apply the Aesthetic Predictor [16], the OpenCV [6], the DOVER [101], and the OCR model [90] to obtain aesthetic scores, motion scores, technical scores, and watermark-free video areas, respectively, and filter out low-quality data, ultimately yielding 5,437,544 high-quality clips.

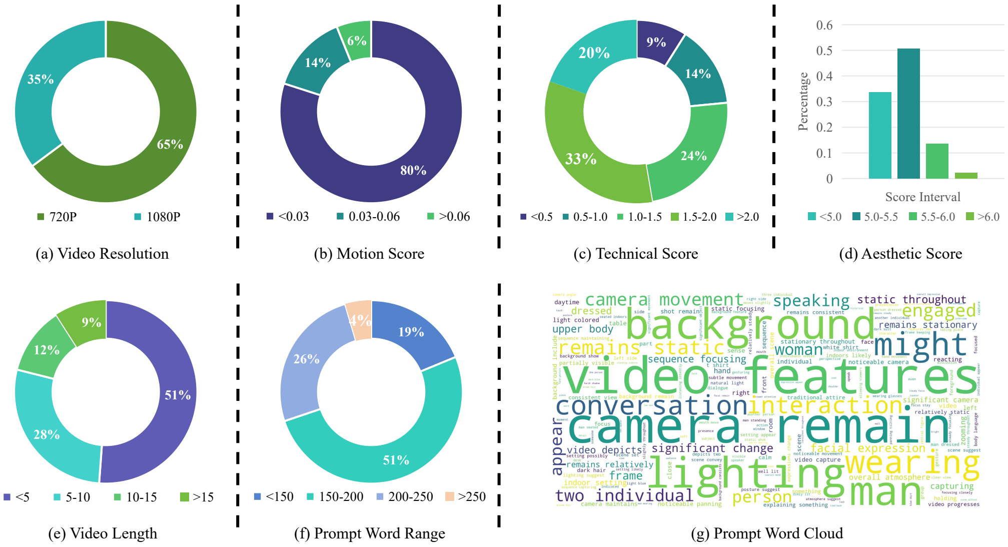
Subject-Driven Annotation. Unlike text-to-video, subject-to-video data requires captions that emphasize the subject. To achieve this, we first use Qwen2.5-VL-7B [92] to describe the appearance and changes of the subject while preserving essential elements of the video, such as environmental context and camera movements, to get the subject-centric video caption. Next, to obtain high-quality reference images, we use DeepSeekV3 [55] to extract keywords related to the environment and objects from the caption. We then input the first frame of the video and these keywords into GroundingDino [58], an open-vocabulary object detection algorithm, to extract reference images for each video. Finally, the bounding boxes obtained from the previous step are fed into SAM2.1 [78], which generates a mask for each subject. This mask can be used to extract reference images without background pixels. To ensure data quality, we further assign Aesthetic Score [16] and text GmeScore to the reference images, allowing users to adjust thresholds to balance data quantity and quality.

C.2 Additional Details of Dataset Statistics

OpenS2V-5M is the first high-quality, large-scale S2V dataset. In contrast to standard datasets [46, 9, 12], it includes Nexus Data specifically designed to address three critical challenges faced by S2V methods. As depicted in Figure 15, the word cloud illustrates the dataset's rich visual content. Regarding video duration, the majority (91%) of videos are between 0 and 10 seconds, while the remaining videos exceed 10 seconds. In terms of resolution, 65% are 720P, with the rest being high-resolution videos. The captions primarily consist of detailed descriptions, with a wide range of word usage. These settings are tailored to the emerging DiT-based models [61, 42, 111, 91], which favor long prompts and are constrained by input limitations, such as 81 frames and 480P resolution. Furthermore, low-quality videos were excluded during preprocessing based on motion, technical, and aesthetic scores, ensuring that most videos are of high quality. Due to resource constraints, we select the top 10K samples with the highest average scores from the 5M dataset to construct gpt-frame pairs. For cross-frame pairs, we identify 0.35M clustering centers from the regular data, each containing an average of 10.13 samples, meaning we could theoretically create far more than 0.35M × 10.13 pairs.

C.3 Further Verification on OpenS2V-5M

Due to limited space in the main text, we provide additional qualitative analysis of Ours‡ here, with results shown in Figure 16. It can be observed that Ours‡ is capable of generating high-quality videos, thereby validating the effectiveness of the proposed OpenS2V-5M.



(a) Video Resolution

(b) Motion Score

(c) Technical Score

(d) Aesthetic Score

(e) Video Length

(f) Prompt Word Range

(g) Prompt Word Cloud

Figure 15: **Statistics in OpenS2V-5M.** The dataset includes a diverse range of categories, clip durations and caption lengths, with most of videos being in high quality (e.g., resolution, aesthetic).

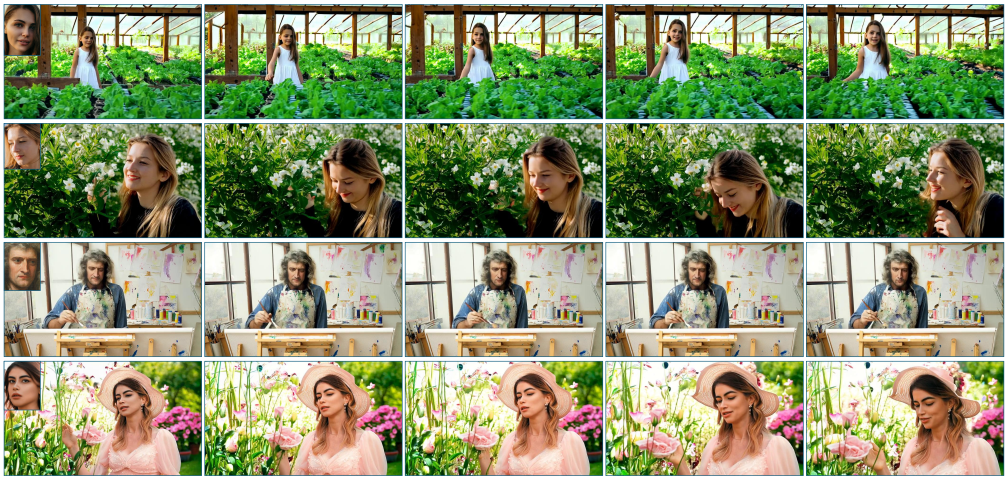


Figure 16: **More Showcases Generated by Ours‡.**

C.4 Samples of Collected Data

Figure 17 presents diverse samples from the OpenS2V-5M dataset, which consists of subject-text-video triples across multiple categories, offering rich visual information. The subjects include both regular data obtained through segmentation and Nexus Data generated via cross-video association and GPT-Image-1, encompassing humans, objects, backgrounds, and more. These samples highlight the dataset's diversity and depth, and are expected to address the three primary challenges faced by subject-to-video generation models, thereby advancing the field and contributingto the community.

D More Details of Experiment

D.1 Details of Resource

We employ Nvidia A100 (x40) for all the experiments. All implementations are conducted on the basis of the official code using the PyTorch framework or official interface.

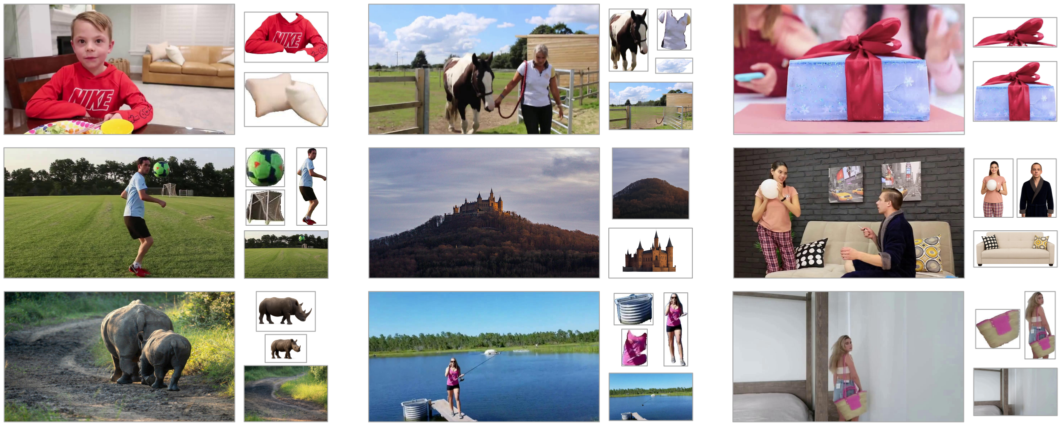


Figure 17: **Samples from the OpenS2V-5M dataset.** The dataset consists of subject-text-video triples, which exhibit more physical knowledge than existing large-scale T2V dataset [12, 93].

D.2 Details of Evaluation Models

As most S2V models [115, 125, 18, 57, 22, 41] do not support dynamic resolution or variable duration, standardization of these parameters is infeasible. Therefore, we adopt the commonly used official settings [60, 37, 83, 103] to maintain fairness across comparisons.

Vidu *Model Details.* Vidu [5] has released three versions of closed-source models: 1.0, 1.5, and 2.0. Among these, versions 1.5 and 2.0 support multi-reference image input, enabling open-domain subject-to-video generation. However, as the technical report has not been published, specific implementation details remain undisclosed. *Implementation Setups.* We employ the official Vidu 2.0 *charactertovideo* feature with default parameter settings. Using the turbo mode, we generate a 4-second video (65-frames) with a spatial resolution of 704 × 396, automatic motion amplitude, and a frame rate of 16 fps.

Pika *Model Details.* Pika [45] has developed five iterations of closed-source model, designated as versions 1.0, 1.5, 2.0, 2.1, and 2.2. Notably, versions 2.0, 2.1 and 2.2 incorporate multi-reference image input capability, enabling open-domain subject-to-video generation. However, due to the absence of an official technical report, the underlying implementation details remain undisclosed. *Implementation Setups.* We employ the official Pika 2.1 *pikaadditions* feature with default parameter settings. The generated video maintains a resolution of 1920 × 1080 pixels and a frame rate of 24 fps, with a total duration of 5 seconds (121-frames).

Kling *Model Details.* Kling [44] has released five versions of closed-source model: 1.0, 1.6, and 2.0, among which version 1.6 supports the input of multiple reference images for open-domain subject-to-video generation. However, as no technical report has been released for this version, we are unable to obtain further details. *Implementation Setups.* We employ the official Kling 1.6 *multi*|*-id* feature with default parameter settings. Using the standard mode, we generate a 5-second video (153-frames) with a spatial resolution of 1280 × 720, and a frame rate of 30 fps.

Hailuo *Model Details.* Hailuo [89] has released six versions of closed-source model: I2V-01-Director, I2V-01-live, I2V-01, T2V-01-Director, T2V-01, and S2V-01. Among them, S2V-01 supports the input of multiple reference images to achieve human-domain subject-to-video generation. However, since no technical report has been released for this model, we are unable to obtain further details. *Implementation Setups.* We use the S2V function of the official Hailuo-S2V-01, available at Hailuo-S2V-01, and keep the default settings. We generate a 5-second video (141-frames) with a spatial resolution of 1280 × 720 and a frame rate of 25fps.

VACE *Model Details.* VACE [41] is a video generation model based on DiT that integrates various inputs in four data modalities—text, image, video, and mask—and unifies multiple video generation and editing tasks within a single model, including open-domain subject-to-video generation. It releases four model weights: VACE-Wan2.1-1.3B-Preview, VACE-LTX-Video-0.9, Wan2.1-VACE-1.3B, and Wan2.1-VACE-14B. The training data consists of over a million text-to-video samples, which it collects and processes internally. *Implementation Setups.* We use the officially released

VACE code and models, maintaining the original settings. For VACE-Wan2.1-1.3B-Preview and VACE-Wan2.1-1.3B, we generate 5-second (81-frame) videos at a spatial resolution of 832×480 and a frame rate of 16 fps. For VACE-Wan2.1-14B, we generate 5-second (81-frame) videos at a spatial resolution of 1280 × 720 and a frame rate of 16 fps.

Phantom *Model Details.* Phantom [57] is a video generation model based on DiT that extracts reference image information using both CLIP and VAE, and employs a windowed attention mechanism to reduce computational overhead, enabling open-domain subject-to-video generation. It includes three model weights: Phantom-Seaweed, Phantom-Wan-1.3B, and Phantom-Wan-14, but only Phantom-Wan-1.3B&14B are publicly released. The training data come from panda70M [12], subject200k [14], OmniGen [106], and internal datasets, totaling over 10 million samples. *Implementation Setups.* We use the officially released Phantom-Wan code and model, maintaining the original settings. We generate 5-second (81-frame) videos at a resolution of 832 × 480 and a 16 fps.

SkyReels-A2 *Model Details.* SkyReels-A2 [22] is a model fine-tuned based on Wan2.1 [91], employing an approach similar to Phantom. It utilizes a dual-stream architecture to enhance the model's response to reference images and textual prompts, enabling open-domain subject-to-video generation. There are four variants in total: A2-Wan2.1-14B-Preview, A2-Wan2.1-14B, A2-Wan2.1-14B-Pro, and A2-Wan2.1-14B-Infinity, but only A2-Wan2.1-14B-Preview has been open-sourced. The training data comes from 2 million high-quality subject-text-video triples collected internally. *Implementation Setups.* We use the officially released SkyReels-A2-Wan2.1-14B-Preview code and model, maintaining the original settings. Videos are generated with a spatial resolution of 832 × 480 and a frame rate of 16 fps, resulting in a duration of 5 seconds (81 frames).

HunyuanCustom *Model Details.* HunyuanCustom [34] is a model fine-tuned based on HunyuanVideo [34], which achieves open-domain subject-to-video generation by injecting ID information into both the MLLM and the video-driven injection module. In theory, it supports the input of multiple reference images, but currently only the weights supporting Single-Subject have been open-sourced. The training data is processed from internally collected and open-source datasets, but the size of the dataset has not been disclosed. *Implementation Setups.* We use the officially released HunyuanCustom-Single-Subject code, maintaining the original settings. Videos are generated with a spatial resolution of 1280 × 720 and a 25 fps, resulting in a duration of 5 seconds (129 frames).

ConsisID *Model Details.* ConsisID [115] is a model fine-tuned based on CogVideoX [111], which achieves human-domain subject-to-video generation by decomposing ID information into high- and low-frequency signals and injecting them into DiT via cross-attention. It only supports the input of a single face image. The training data is processed from internally collected data, with a dataset size of approximately 0.1 million. *Implementation Setups.* We use the officially released ConsisID code and model, maintaining the original settings. Videos are generated with a spatial resolution of 720 × 480 and a frame rate of 8 fps, resulting in a duration of 6 seconds (49 frames).

Concat-ID *Model Details.* Concat-ID [125] is a model fine-tuned based on CogVideoX [115] and Wan2.1 [91]. It concatenates image features with video latents along the token dimension, thereby avoiding the issue of blurry initial frames. It only supports input of a single face image. The training data is processed from internally collected data, with a dataset size of approximately 1.3 million. *Implementation Setups.* We use the officially released Concat-ID code and model, maintaining the original settings. For CogVideoX version, videos are generated with a spatial resolution of 720 × 480 and a frame rate of 8 fps, resulting in a duration of 6 seconds (49 frames). For Wan-AdaLN version, videos are generated with a spatial resolution of 832 × 480 and a frame rate of 16 fps, resulting in a duration of 5 seconds (81 frames).

FantasyID *Model Details.* FantasyID [122] is a model fine-tuned from CogVideoX [111] that facilitates identity-consistent generation by constructing multi-view facial datasets, incorporating 3D geometric priors, and utilizing a layer-aware control signal injection mechanism. The model currently supports only single face image input. Its training data are drawn from ConsisID [115], CelebV-HQ [127], and Open-vid [69], comprising approximately 50,000 samples. *Implementation Setups.* We employ the officially released Fantasy-ID code and model while retaining the original settings. Videos are generated at a spatial resolution of 720 × 480 and a frame rate of 8 fps, yielding a duration of 6 seconds (49 frames).

EchoVideo *Model Details.* EchoVideo [99] is a model fine-tuned from CogVideoX [111] that employs the multimodal feature fusion module IITF to achieve identity-preserving video generation through the integration of textual, visual, and facial identity information. The model supports only a single face image as input. The training data are sourced from internal collections and comprise approximately 3.3 million samples. *Implementation Setups.* We employ the officially released EchoVideo code and model while retaining the original settings. Videos are generated at a spatial resolution of 848×480 and a frame rate of 16 fps, yielding a duration of 3 seconds (49 frames).

VideoMaker *Model Details.* VideoMaker [105] is a UNet-based model fine-tuned from AnimateDiff [26]. It directly inputs reference images into the video diffusion model and utilizes its intrinsic feature extraction process to achieve subject-to-video generation (e.g., only supports 10 categories of subjects). The training data are sourced from CelebV-Text [127] and VideoBooth [40], comprising approximately 0.1M samples. *Implementation Setups.* We employ the officially released VideoMaker code and model while retaining the original settings. Videos are generated at a spatial resolution of 512×512 and a frame rate of 8 fps, yielding a duration of 2 seconds (16 frames).

ID-Animator *Model Details.* ID-Animator [30] is a UNet-based model fine-tuned from AnimateDiff [26] that employs FaceAdapter and cross-attention to inject facial information. The model supports only a single face image as input. The training data are sourced from CelebV-Text [127] and comprise approximately 15K samples. *Implementation Setups.* We employ the officially released ID-Animator code and model while retaining the original settings. Videos are generated at a spatial resolution of 512×512 and a frame rate of 8 fps, yielding a duration of 2 seconds (16 frames).

D.3 Additional Details of Evaluation Settings

Because some models support only a single subject, while others support multiple subjects, we categorize the evaluation tasks into the following three groups:

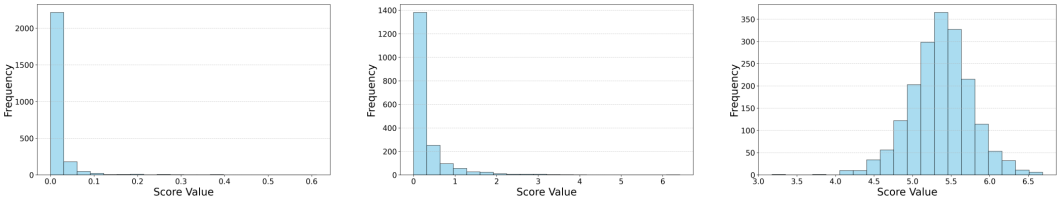
Open-Domain Subject-to-Video including ① single-face-to-video, ② single-body-to-video, ③ single-entity-to-video, ④ multi-face-to-video, ⑤ multi-body-to-video, ⑥ multi-entity-to-video, and ⑦ human-entity-to-video.

Human-Domain Subject-to-Video including ① single-face-to-video and ② single-body-to-video. In this context, only the face image is input, without the body image.

Single-Domain Subject-to-Video including ① single-face-to-video, ② single-body-to-video, and ③ single-entity-to-video.

D.4 Additional Details of Implementations

With the exception of Motion Amplitude and Motion Smoothness, which requires the use of all frames, the other metrics (e.g., NexusScore, NaturalScore, GmeScore, FaceSim, AestheticScore) are calculated by uniformly sampling 32 frames to ensure fairness and minimize overhead. Additionally, due to the differing optimal inference settings for each model, it is not feasible to standardize the resolution of generated videos. **(1)** For Motion Amplitude, we use OpenCV [6] to compute this using the *OpticalFlowFarneback*. **(2)** For Motion Smoothness, we use QAlignVideoScore [54] to compute the motion smoothness about the video. **(2)** For FaceSim, following the approach outlined in ConsisID [125], we first apply insightface [17] to detect the face regions in the video frames and the reference image. We then calculate the similarity between these regions in the curricularface [35] feature space. Finally, we average the sum of all valid scores to obtain the FaceSim for the video. **(3)** For AestheticScore, following the method presented in the improved-aesthetic-predictor [16], we directly input the video frames into the model to obtain scores, then compute the average of all valid scores to obtain the AestheticScore for the video. **(4)** For NexusScore, since we have filtered out low-quality $B_{i,t}$ using $c_{i,t}$ and $s_{i,t}$, high-quality scores may be obtained when only one frame of the video is of high quality while the remaining frames are of lower quality. Therefore, after summing and averaging all valid scores, we divide by T' to mitigate this issue. Here, T' refers to the total number of frames in which an object is detected. In addition, this metric is not used to calculate face similarity to improve robustness, which is why we retain FaceSim. **(5)** For NaturalScore, we use *gpt-4o-2024-11-20* [1] as the base model. For each video, we resize the longer side to 512 pixels and run the model three times, taking the average of these results as the score for the video. **(6)** For GmeScore, since it is based on Qwen2-VL [92], which natively supports dynamic resolution and variable duration, no special processing is necessary.



(a) Distribution of NexusScore

(b) Distribution of MotionScore

(c) Distribution of AestheticScore

Figure 18: **Distribution of NexusScore, AestheticsScore and Motion-A.**

D.5 Additional Details of Metircs Normalization

OpenS2V-Eval evaluates six key dimensions: *subject consistency*, *subject naturalness*, *text relevance*, *face similarity*, *visual quality*, and *motion amplitude*. Due to differing units of measurement across these metrics, direct comparisons and comprehensive analysis are infeasible without normalization. To resolve this, we normalize each metric by defining its theoretical or empirical bounds:

- **FaceSim-Cur**, **GmeScore** and **Motion-S** are bounded by construction, with ranges at $[0, 1]$.
- **NaturalScore** employs a 5-point Likert scale, spanning $[1, 5]$.
- For unbounded metrics (**NexusScore**, **AestheticScore**, and **Motion-A**), we derive ranges of $[0, 0.05]$, $[0, 1]$, and $[4, 7]$, respectively, from their empirical distributions (Figure 18). Out-of-range values are truncated.

To aggregate these normalized metrics into a unified performance score, we compute a weighted sum:

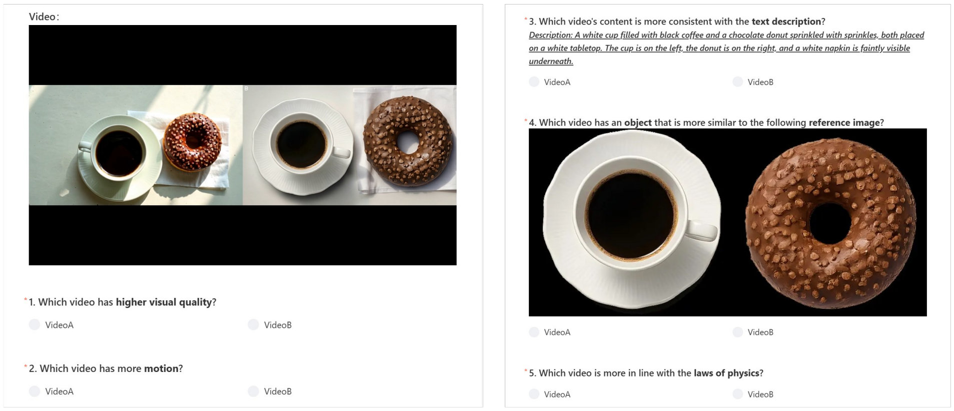
$$\text{Total\_Score} = \sum_{i \in \mathcal{M}} w_i \cdot S_i, \quad \text{where } \mathcal{M} = \{\text{Nexus}, \text{Natural}, \text{Gme}, \text{FaceSim}, \text{Aesthetic}, \text{Motion}\}, \tag{7}$$

with weights w_i assigned as $\iota = 0.20$ (NexusScore), $\kappa = 0.24$ (NaturalScore), $\lambda = 0.12$ (GmeScore), $\mu = 0.20$ (FaceSim-Cur), $\nu = 0.16$ (AestheticScore), $\xi = 0.02$ (Motion-A) and $\sigma = 0.06$ (Motion-S). For humam-domain S2V task, $\kappa = 0.30$, $\lambda = 0.15$, $\mu = 0.25$, $\nu = 0.18$, $\xi = 0.03$ and $\sigma = 0.09$.

D.6 Additional Details of Human Evaluation

Pre-processing The questionnaire for human evaluation of generated content is developed based on prior studies [115, 117, 77, 83, 82], as shown in Figure 19. The evaluation focuses on six key aspects: *subject consistency*, *subject naturalness*, *text relevance*, *face similarity*, *visual quality*, and *motion amplitude*. For each criterion, a pairwise comparison method is employed, allowing participants to choose between two video options, thereby improving user pleasure and increasing the number of effective questionnaire samples. To ensure category balance, 30 test samples are randomly selected from OpenS2V-Eval, with each sample paired with two videos generated by different models, yielding a total of 60 videos. These videos are annotated with six evaluation scores: NexusScore, NaturalScore, GmeScore, FaceSim-Cur [115], AestheticScore [16], and Motion Quality (Amplitude [6], Smoothness [54]). Taking subject consistency as an example, a sample is labeled as a positive instance for NexusScore if a participant prefers video A over video B and A's NexusScore exceeds that of B; otherwise, it is labeled as a negative instance. The final human preference ratio for each metric is computed as the proportion of positive instances among all test samples. Participants include undergraduate, master's, and doctoral students, as well as members of the general public with no direct affiliation to the research domain. They are drawn from a diverse international pool, including individuals from China, and the United States. This heterogeneous composition ensures both the reliability and generalizability of the evaluation results.

Post-processing Folloing [115, 117, 116], to ensure data quality given the use of a five-point evaluation scale, we exclude outlier responses through the following procedures: ① We limit each submission to a single response per IP address and require users to log in prior to voting, thereby ensuring that each participant can submit only one response. ② We assess data validity by considering questionnaire completion time. As it requires 5 to 10 minutes to complete the survey, we exclude responses submitted in less than 5 minutes. ③ We randomize the playback order of videos for each



Video:

*1. Which video has **higher visual quality**?

○ VideoA ○ VideoB

*2. Which video has more **motion**?

○ VideoA ○ VideoB

*3. Which video's content is more consistent with the **text description**?

Description: A white cup filled with black coffee and a chocolate donut sprinkled with sprinkles, both placed on a white tabletop. The cup is on the left, the donut is on the right, and a white napkin is faintly visible underneath.

○ VideoA ○ VideoB

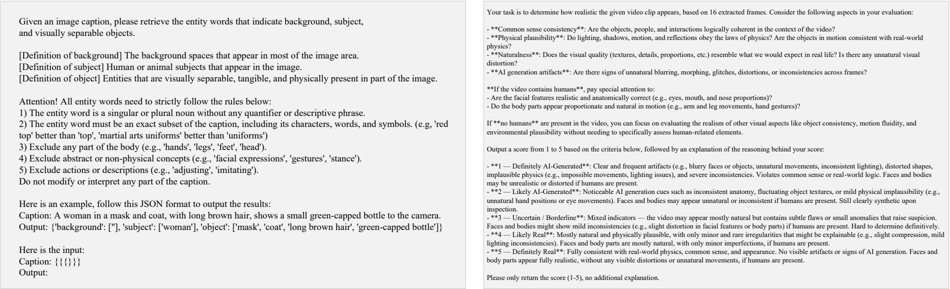
*4. Which video has an **object** that is more similar to the following **reference image**?

○ VideoA ○ VideoB

*5. Which video is more in line with the **laws of physics**?

○ VideoA ○ VideoB

Figure 19: **Visualization of the Questionnaire for User Study.**



```
Given an image caption, please retrieve the entity words that indicate background, subject,
and visually separable objects.

[Definition of background] The background spaces that appear in most of the image area.
[Definition of subject] Human or animal subjects that appear in the image.
[Definition of object] Entities that are visually separable, tangible, and physically present in part of the image.

Attention! All entity words need to strictly follow the rules below:
1) The entity word is a singular or plural noun without any quantifier or descriptive phrase.
2) The entity word must be an exact subset of the caption, including its characters, words, and symbols. (e.g, 'red
top' better than 'top', 'martial arts uniforms' better than 'uniforms')
3) Exclude any part of the body (e.g., 'hands', 'legs', 'feet', 'head').
4) Exclude abstract or non-physical concepts (e.g., 'facial expressions', 'gestures', 'stance').
5) Exclude actions or descriptions (e.g., 'adjusting', 'imitating').
Do not modify or interpret any part of the caption.

Here is an example, follow this JSON format to output the results:
Caption: A woman in a mask and coat, with long brown hair, shows a small green-capped bottle to the camera.
Output: {'background': [''], 'subject': ['woman'], 'object': ['mask', 'coat', 'long brown hair', 'green-capped bottle']}

Here is the input:
Caption: {{{}}}
Output:
```

(a) Prompt for Extracting Tags

```
Your task is to determine how realistic the given video clip appears, based on 16 extracted frames. Consider the following aspects in your evaluation:

- **Common sense consistency**: Are the objects, people, and interactions logically coherent in the context of the video?
- **Physical plausibility**: Do lighting, shadows, motion, and reflections obey the laws of physics? Are the objects in motion consistent with real-world
physics?
- **Naturalness**: Does the visual quality (textures, details, proportions, etc.) resemble what we would expect in real life? Is there any unnatural visual
distortion?
- **AI generation artifacts**: Are there signs of unnatural blurring, morphing, glitches, distortions, or inconsistencies across frames?

**If the video contains humans**, pay special attention to:
- Are the facial features realistic and anatomically correct (e.g., eyes, mouth, and nose proportions)?
- Do the body parts appear proportionate and natural in motion (e.g., arm and leg movements, hand gestures)?

If **no humans** are present in the video, you can focus on evaluating the realism of other visual aspects like object consistency, motion fluidity, and
environmental plausibility without needing to specifically assess human-related elements.

Output a score from 1 to 5 based on the criteria below, followed by an explanation of the reasoning behind your score:

- **1 — Definitely AI-Generated**: Clear and frequent artifacts (e.g., blurry faces or objects, unnatural movements, inconsistent lighting), distorted shapes,
implausible physics (e.g., impossible movements, lighting issues), and severe inconsistencies. Violates common sense or real-world logic. Faces and bodies
may be unrealistic or distorted if humans are present.
- **2 — Likely AI-Generated**: Noticeable AI generation cues such as inconsistent anatomy, fluctuating object textures, or mild physical implausibility (e.g.,
unnatural hand positions or eye movements). Faces and bodies may appear unnatural or inconsistent if humans are present. Still clearly synthetic upon
inspection.
- **3 — Uncertain / Borderline**: Mixed indicators — the video may appear mostly natural but contains subtle flaws or small anomalies that raise suspicion.
Faces and bodies might show mild inconsistencies (e.g., slight distortion in facial features or body parts) if humans are present. Hard to determine definitively.
- **4 — Likely Real**: Mostly natural and physically plausible, with only minor and rare irregularities that might be explainable (e.g., slight compression, mild
lighting inconsistencies). Faces and body parts are mostly natural, with only minor imperfections, if humans are present.
- **5 — Definitely Real**: Fully consistent with real-world physics, common sense, and appearance. No visible artifacts or signs of AI generation. Faces and
body parts appear fully realistic, without any visible distortions or unnatural movements, if humans are present.

Please only return the score (1-5), no additional explanation.
```

(b) Prompt for Getting NaturalScore

Figure 20: **Visualization of Different Input Text Prompts.**

participant to mitigate cognitive bias. ④ We implement a sliding verification upon submission to ensure that all questionnaires are completed manually, thereby preventing automated (bot) responses. ⑤ We exclude any questionnaires for which more than 50% of evaluations are extreme values, defined as responses where the sum of the highest (5) and lowest (1) ratings exceeds 50%.

D.7 Additional Details of Input Prompts

Regarding how to obtain tags through Deepseek [55] and how to annotate videos with NaturalScore using GPT-4o [1], we visualize the input text prompt, as shown in Figure 20.

E Additional Statement

E.1 Limitations and Future Work

Although NexusScore and NaturalScore are introduced to evaluate subject consistency and naturalness, these metrics show only approximately 75% correlation with human preferences. Future work aims to better align automated metrics with human judgments. The videos in OpenS2V-5M come from multiple video platforms, and we can only make publicly available those that comply with the CC BY 4.0 license or are copyright-free, totaling approximately 4 million videos.

E.2 Declaration of LLM Usage

We utilized Large Language Models (LLMs), such as ChatGPT, to support the preparation of this paper. Specifically, LLMs were employed for language-related tasks, including grammar correction, spelling checks, and word choice refinement, to improve the manuscript's clarity and fluency. Additionally, LLMs assisted with data processing and filtering (e.g., our NaturalScore is GPT-based), as well as generating draft figures to assist the authors in creating refined visualizations. All scientific content, analyses, and conclusions were independently conceived, validated, and interpreted by the authors.

E.3 Potential Harms Caused by the Research Process

The subject images of **OpenS2V-Eval** are derived from three open-source datasets—ConsisID [115], A2-Bench [22], and DreamBench [73]—that adhere to the Apache license, as well as from three video platforms—Pexels, MixKit, and PixaBay—that operate under the Creative Commons Zero (CC0) license. The video data in **OpenS2V-5M** originates from the Open-Sora Plan [51], with some content licensed under Creative Commons Attribution 4.0 (CC BY 4.0) and others under the Royalty-Free (RF) license. The licensing information for these data is explicitly stated on their respective platforms. The CC0 license designates content as public domain, permitting unrestricted use without additional permissions or authorizations. For CC BY 4.0-licensed videos from the Open-Sora Plan [51], video IDs are included in the metadata to mitigate potential contractual disputes. For RF-licensed videos, we are working to resolve intellectual property issues. In total, approximately 4 million data will be made available as open source. The collected data is organized into seven categories, with contributions from global sources. This diversity ensures that OpenS2V-Eval and OpenS2V-5M are fully representative. The ConsisID model [115] fine-tuned on our dataset demonstrated no significant content bias. Furthermore, video content has been filtered to exclude NSFW material based on subtitle detection. Due to the presence of videos containing identifiable individuals, access to OpenS2V-Nexus is restricted to academic use only, with contact information provided on the https://pku-yuangroup.github.io/OpenS2V-Nexus to ensure the security of personal identity data.

Data collection was made possible through the dedicated efforts of numerous contributors, including the authors of this paper and those involved in the manual evaluation. We consider individual hourly wages or compensation as personal information, and for privacy reasons, these details cannot be disclosed. Nonetheless, we can confirm that all participants have received appropriate compensation in accordance with the legal requirements of their respective countries or regions. The privacy of all participants is safeguarded, ensuring that no additional risks are posed to them.

E.4 Societal Impact and Potential Harmful Consequences

The objective of **OpenS2V-Eval** is to identify the limitations of existing subject-to-video generation models and to develop the **OpenS2V-5M** dataset to further advance research in this area. While subject-to-video generation models hold significant potential for enhancing creativity, their broader societal impacts must be carefully considered during development:

First, environmental resource consumption. Training subject-to-video generation models requires extensive GPU computing power, with a single large-scale training session potentially consuming tens of thousands of kilowatt-hours of electricity, resulting in carbon emissions comparable to the annual emissions of several dozen cars. This high energy consumption not only exacerbates global climate change but also consolidates computational resources within a few dominant tech companies, exacerbating inequality in the research community. To address this, efforts should focus on exploring techniques for model lightweighting, optimizing distributed training efficiency, and promoting the development of green data centers powered by renewable energy to reduce the carbon footprint.

Second, the risk of linguistic homogeneity and cultural bias. The text prompt in OpenS2V-Nexus are currently limited to English, which may introduce bias in the model's interpretation of multilingual contexts, such as Chinese. For instance, when generating videos involving non-Western cultural symbols (e.g., Hanfu, Kung fu), the lack of relevant training data could lead to semantic distortions or cultural misinterpretations. Solutions include creating a multilingual annotation system and establishing an open-source collaborative framework to encourage researchers globally to contribute localized data, helping bridge language barriers.

Finally, the ethical concerns associated with deepfake misuse. Subject-consistency video generation technologies may be exploited for malicious purposes, such as creating political misinformation, forging celebrity images, or fabricating criminal evidence. The level of realism achievable with these technologies surpasses that of traditional Photoshop techniques. Such misuse poses a threat to public opinion security and judicial integrity. Effective countermeasures should combine technological governance and regulatory oversight: developing generative models embedded with imperceptible watermarks, establishing blockchain-based content traceability protocols, and advocating for legislation requiring mandatory labeling of generated content. Additionally, public media literacy campaigns should be implemented to enhance society's resilience to false information.

E.5 Impact Mitigation Measures

We are fully responsible for the authorization, distribution, and maintenance of **OpenS2V-Eval** and **OpenS2V-5M**. Our datasets and benchmarks are released under the CC-BY-4.0 license, while the code is released under the Apache license. We explicitly state on our homepage that all data is intended for academic research purposes to prevent misuse or improper use. We also provide metadata for each video, allowing video creators to contact us promptly and remove invalid videos. All metadata is hosted on *GitHub* and *HuggingFace*, with the following links: https://github.com/PKU-YuanGroup/OpenS2V-Nexus and https://huggingface.co/collections/BestWishYsh.

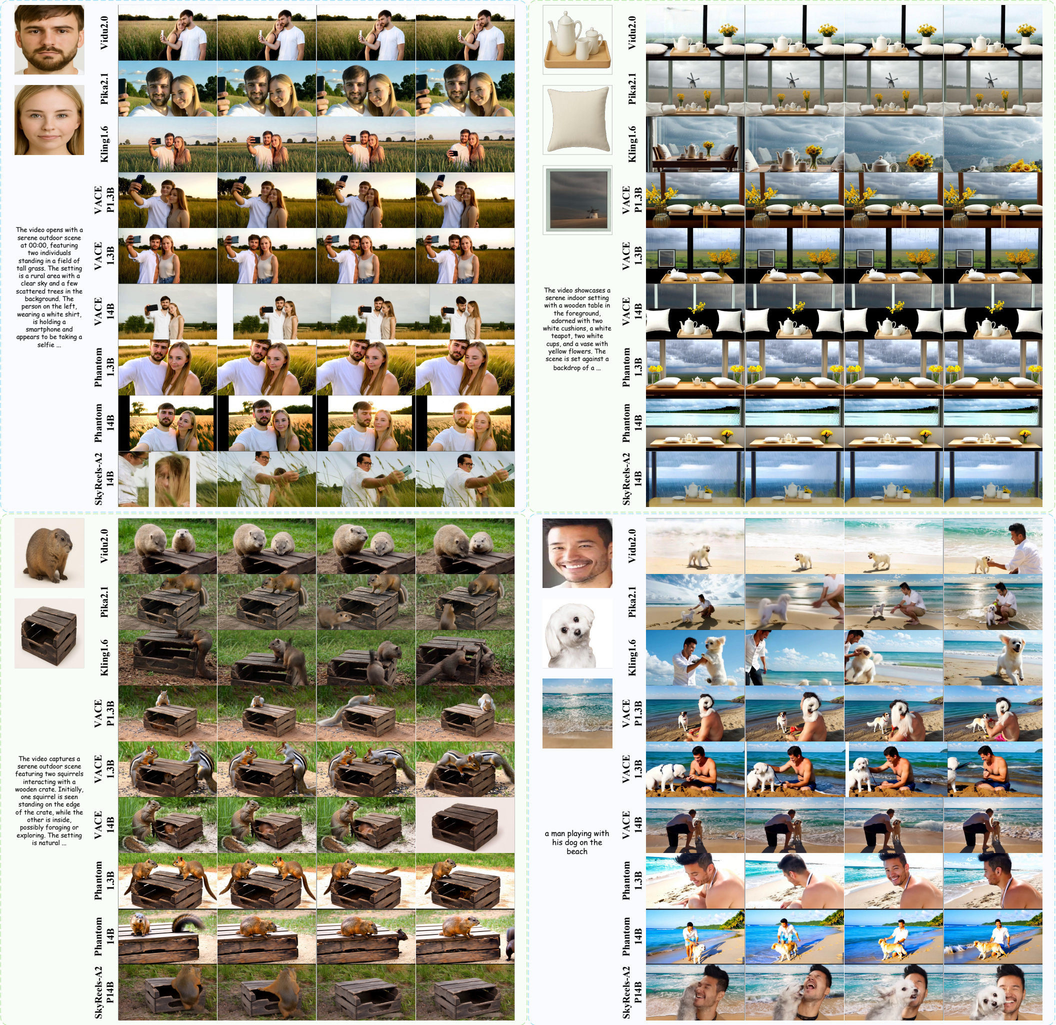

Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 14B
The video opens with a serene outdoor scene at 00:00, featuring two individuals standing in a field of tall grass. The setting is a rural area with a clear sky and a few scattered trees in the background. The person on the left, wearing a white shirt, is holding a smartphone and appears to be taking a selfie ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 14B
The video showcases a serene indoor setting with a wooden table in the foreground, adorned with two white cushions, a white teapot, two white cups, and a vase with yellow flowers. The scene is set against a backdrop of a ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
The video captures a serene outdoor scene featuring two squirrels interacting with a wooden crate. Initially, one squirrel is seen standing on the edge of the crate, while the other is inside, possibly foraging or exploring. The setting is natural ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
a man playing with his dog on the beach

Figure 21: **More Showcases in OpenS2V-Eval for Open-Domain Subject-to-Video Generation.**

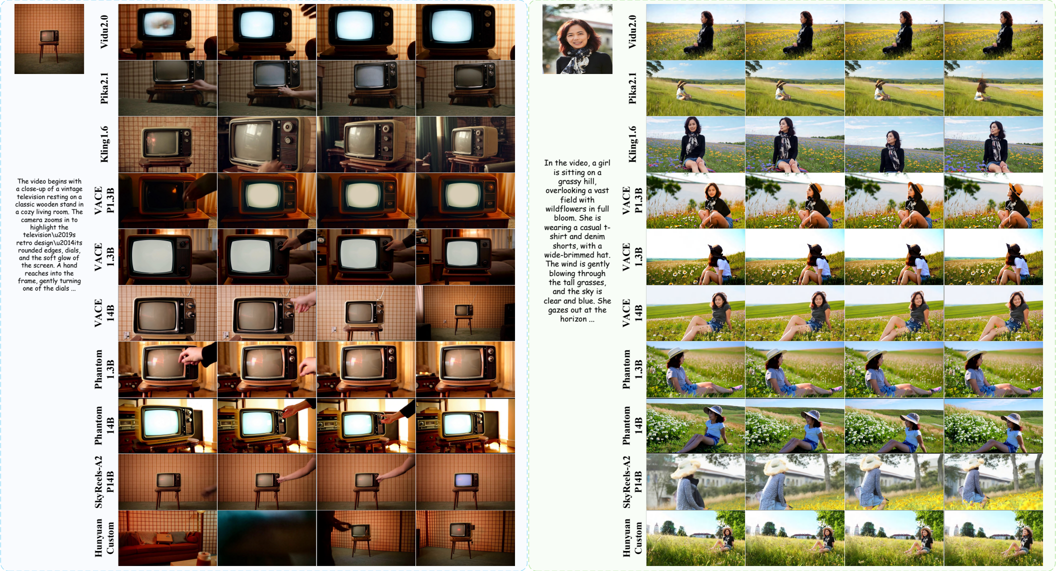

Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
Hunyuan Custom
The video begins with a close-up of a vintage television resting on a classic wooden stand in a cozy living room. The camera zooms in to highlight the television\u2019s retro design\u2014its rounded edges, dials, and the soft glow of the screen. A hand reaches into the frame, gently turning one of the dials ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
Hunyuan Custom
In the video, a girl is sitting on a grassy hill, overlooking a vast field with wildflowers in full bloom. She is wearing a casual t-shirt and denim shorts, with a wide-brimmed hat. The wind is gently blowing through the tall grasses, and the sky is clear and blue. She gazes out at the horizon ...

Figure 22: **More Showcases in OpenS2V-Eval for Single-Domain Subject-to-Video Generation.**

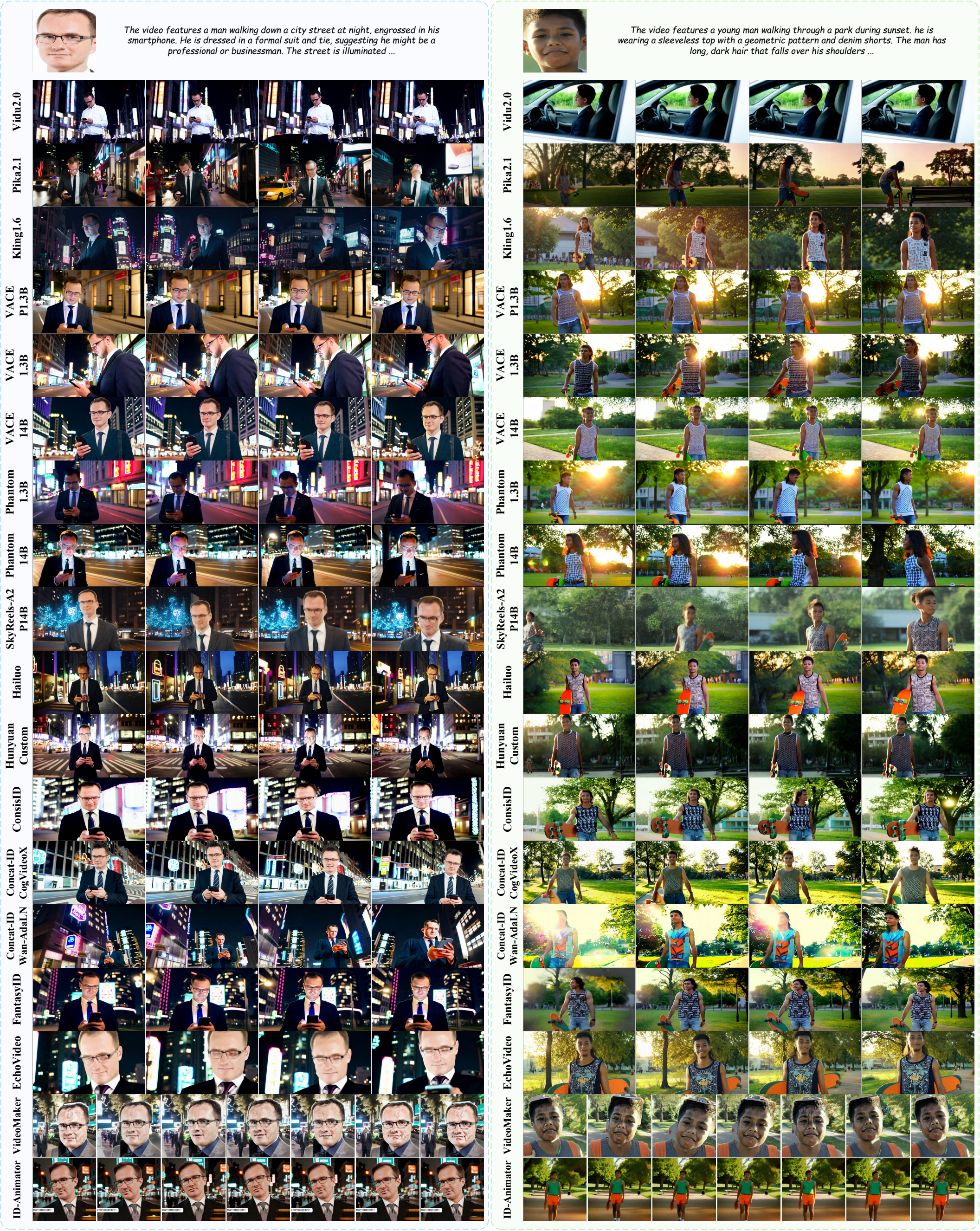

The video features a man walking down a city street at night, engrossed in his smartphone. He is dressed in a formal suit and tie, suggesting he might be a professional or businessman. The street is illuminated ...
The video features a young man walking through a park during sunset. he is wearing a sleeveless top with a geometric pattern and denim shorts. The man has long, dark hair that falls over his shoulders ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
Hailuo
Hunyuan Custom
ConsisID
Concat-ID CogVideoX
Concat-ID Wan-AdaLN
FantasyID
EchoVideo
VideoMaker
ID-Animator

Figure 23: **More Showcases in OpenS2V-Eval for Human-Domain Subject-to-Video Generation.**

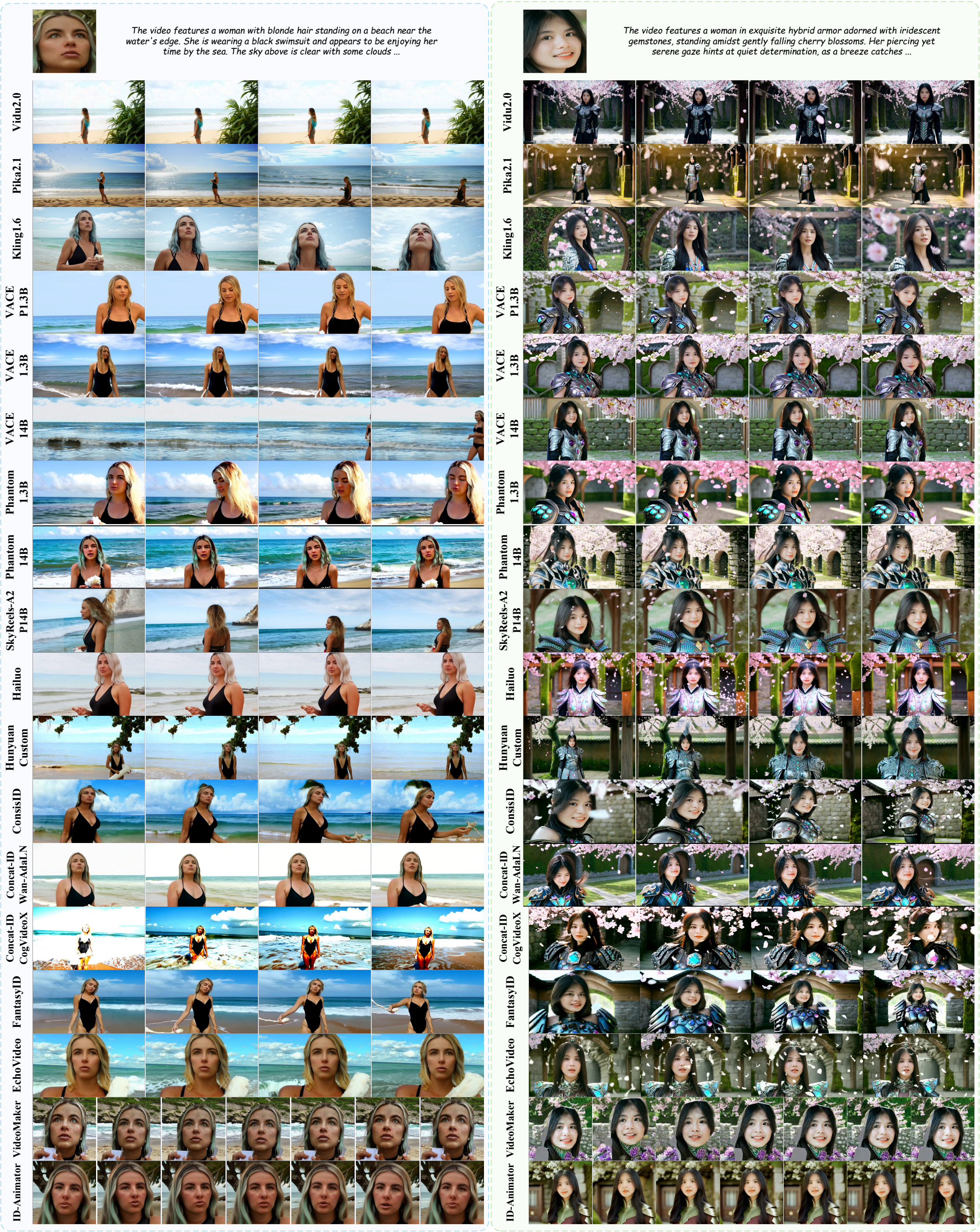

The video features a woman with blonde hair standing on a beach near the water's edge. She is wearing a black swimsuit and appears to be enjoying her time by the sea. The sky above is clear with some clouds ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
Hailuo
Hunyuan Custom
ConsisID
Concat-ID Wan-AdaLN
Concat-ID CogVideoX
FantasyID
EchoVideo
VideoMaker
ID-Animator
The video features a woman in exquisite hybrid armor adorned with iridescent gemstones, standing amidst gently falling cherry blossoms. Her piercing yet serene gaze hints at quiet determination, as a breeze catches ...
Vidu2.0
Pika2.1
Kling1.6
VACE P1.3B
VACE 1.3B
VACE 14B
Phantom 1.3B
Phantom 14B
SkyReels-A2 P14B
Hailuo
Hunyuan Custom
ConsisID
Concat-ID Wan-AdaLN
Concat-ID CogVideoX
FantasyID
EchoVideo
VideoMaker
ID-Animator

Figure 24: **More Showcases in OpenS2V-Eval for Human-Domain Subject-to-Video Generation.**