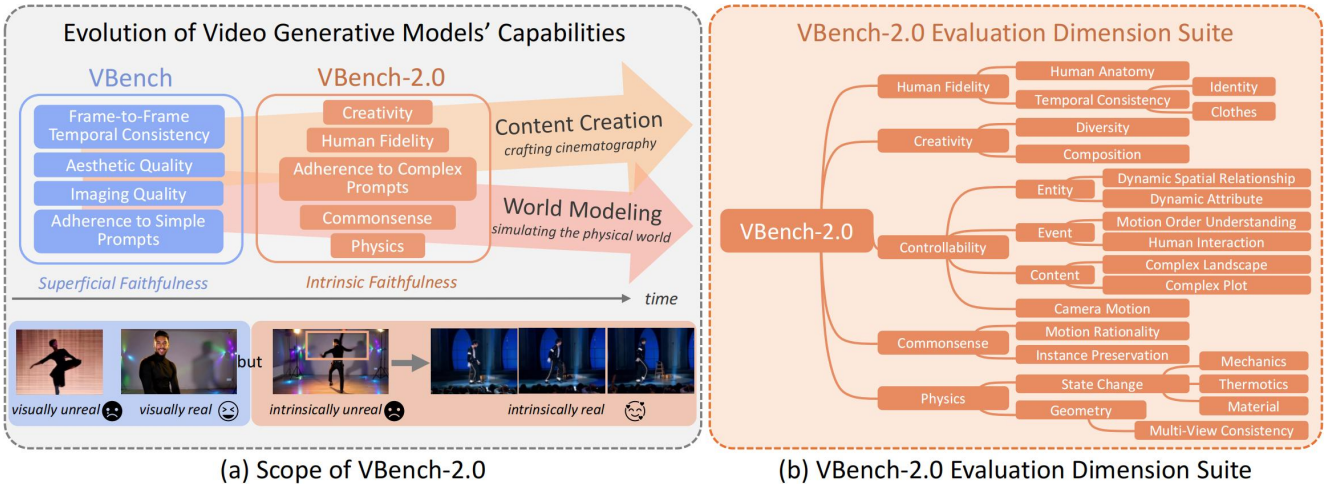


Figure 2. Overview of VBench evaluation metrics [34]. VBench measures visual quality, motion smoothness, identity consistency, temporal flicker, spatial coherence, and text–video relevance to provide a fine-grained, multi-dimensional assessment of generated videos.

# Appendix

## A. Datasets

Web-scale corpora such as Koala-36M [75], WebVid-10M [1], Panda-70M [9] and HD-VG-130M [78] collectively exceed 250 M clips, yet their stock-footage provenance yields noisy captions and licences that forbid commercial use. High-definition, human-centric datasets like CelebV-HQ [105], OpenHumanVid [44], HumanVid [83] and HDTF [102] supply face tracks, skeletons and camera-motion labels, but most clips remain under 20 s, limiting long-form training.

Vript [93] provides six-minute films with 145-word scene-level "scripts." Large-scale video–text datasets like HD-VILA-100M [91] and Panda-70M [9] have enabled text-to-video generation at scale, but their clips are mostly 5–15 s with minimal narrative context. To push toward longer, story-driven videos, recent benchmarks offer richer structure: MiraData provides 1–2 min sequences with dense, structured captions covering objects, actions, style and camera motions [40], and MovieBench is the first movie-level dataset with hierarchical annotations (movie, scene, shot) enforcing character consistency and multi-scene storytelling [86]. Examples of these datasets used in video generation models are in Table 2.

## B. Evaluation Metrics

In recent work, video generation models have largely been assessed using image-derived metrics such as Inception Score (IS) [62], Fréchet Inception Distance (FID) [26] and its temporal extension Fréchet Video Distance (FVD) [69], along with Structural Similarity Index (SSIM) [82] and Learned Perceptual Image Patch Similarity (LPIPS) [99], as well as text alignment via CLIPScore [25]. While these measures offer convenient benchmarks, they obscure crucial factors such as temporal coherence, storytelling fidelity and multi-scene consistency, and frequently diverge from human judgments on longer, more complex clips.

To address these shortcomings, VBench introduces a comprehensive, hierarchical evaluation suite that decomposes "video generation quality" into fine-grained dimensions such as visual quality, motion smoothness, identity consistency, temporal flicker, spatial relationships and text-video relevance. Each of these dimensions are driven by tailored prompt sets and validated with human preference annotations [34]. By providing multi-dimensional scores rather than a monolithic metric, VBench enables detailed diagnostics of generative strengths and weaknesses, making it the principal benchmark for next-generation video models. Another similar benchmark that is multi-dimensional is Wan2.2 [71], however, VBench has been used by most of the recent literature.

Table 2. Overview of video diffusion models, their applications, and training datasets. We categorize models by tasks, generation statistics, subject capabilities, and video duration. (TV) Text-to-Video, (IV) Image-to-Video, (VV) Video-to-Video, (VE) Vide-Extension. [S] Single-Subject, [M] Multi-Subject, [S] Greater than 17 seconds, [S] 5 up to 16 seconds, [S] Less than 5 seconds.

| Paper's Github | Stars | Date | Tasks | | | | Generation Statistics | | | Subjects | Dataset | Affiliation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TV | IV | VV | VE | Len. | FPS | #Frames | | | |
| Seedance 1.0 [20] | - | 06'25 | TV | IV | | | 5 | 30 | 150 | M | - | ByteDance |
| HunyuanVideo-Avatar [12] | 1K | 05'25 | TV | IV | | | 30 | 25 | 750 | M | Koala-36M, CelebV-HQ, HDTF | Tencent |
| MAGI-1 [67] | 3.2K | 05'25 | TV | IV | | VE | 16 | 24 | 384 | S | Open-perfectblend | Sand AI |
| HunyuanCustom [31] | 1K | 05'25 | TV | IV | VV | | 5 | 26 | 129 | M | OpenHumanvid, Panda-2M | Tencent |
| Veo3 [22] | - | 05'25 | TV | IV | | VE | 8 | 60 | 480 | M | - | Google |
| SkyReels-v2 [6] | 2.7K | 04'25 | TV | IV | | VE | 30 | 24 | 720 | M | Koala-36M, HumanVid | Skywork AI |
| Open-Sora 2.0 [54] | 26.6K | 03'25 | TV | IV | | | 5 | 24 | 128 | M | WebVid-10M, Panda-70M, HD-VG-130M, MiraData, Vript, Inter4K | HPC-AI Tech |
| WAN [71] | 11.8K | 03'25 | TV | IV | | VE | 5 | 16 | 81 | M | - | Alibaba |
| VACE [37] | 2.3K | 03'25 | TV | IV | VV | VE | 5 | 16 | 81 | M | - | Alibaba |
| Phantom [48] | 1K | 02'25 | TV | IV | | | 5 | 25 | 125 | M | Panda70M | ByteDance |
| StepVideo [32] | 3K | 02'25 | TV | IV | | | 8 | 25 | 204 | M | - | Step-Video |
| ConceptMaster [33] | - | 01'25 | TV | IV | | | 5 | 62 | 310 | M | Panda-2M, MS-COCO | Kuaishou Tech |
| VideoAlchemist [10] | - | 01'25 | TV | IV | | | 5 | 24 | 120 | M | MSRVTT-Personalization | Snap |
| HunyuanVideo [42] | 10.2K | 12'24 | TV | IV | | VE | 5 | 26 | 129 | M | - | Tencent |
| LTX-video [24] | 6.3K | 12'24 | TV | IV | | VE | 5 | 24 | 120 | M | MS-COCO, LAION-5B, MSR-VTT, UCF-101 | Lightricks |
| MovieGen [55] | - | 10'24 | TV | IV | VV | | 16 | 24 | 384 | L | UCF-101, MSR-VTT, Kinetics-400, SAM-2 | Meta |
| CogVideoX [96] | 11.5K | 08'24 | TV | IV | VV | | 10 | 16 | 160 | L | Panda-70M, COYO-700M, LAION-5B, WebVid | Tsinghua Uni |
| Open-Sora Plan [45] | 12K | 11'24 | TV | IV | | | 6 | 24 | 144 | L | COCO, JourneyDB, Panda70M, VIDAL-10M, WebVid-10M | Peking Uni |

Table 2. (Continued)

| Paper's Github | Stars | Date | Tasks | | | | Generation Statistics | | | Subjects | Dataset | Affiliation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | TV | IV | VV | VE | Len. | FPS | #Frames | | | |
| Loong [81] | - | 10'24 | TV | | | | 150 | 7 | 1050 | M | LAION-5B, MSR-VTT | ByteDance |
| Ca2-VDM [18] | 1K | 06'24 | TV | IV | | | 60 | 24 | 1440 | S | MSR-VTT, UCF-101, Sky Timelapse | Zhejiang Uni |
| Sora [52] | - | 06'24 | TV | IV | | | 20 | 30 | 600 | M | - | openAI |
| StoryDiffusion [104] | 6.3K | 05'24 | TV | | | | 13 | 14 | 182 | M | Webvid10M | ByteDance |
| Lumiere [2] | 1K | 01'24 | TV | IV | VV | | 5 | 16 | 80 | L | UCF101 | Google |
| VideoPoet [64] | - | 12'23 | TV | IV | VV | | 5 | 8 | 41 | L | MSR-VTT, UCF-101, Kinetics 600, Something-Something V2, DAVIS | Google |
| DreamVideo [19] | 3.1K | 12'23 | TV | IV | VV | | 4 | 8 | 32 | S | UCF101, DAVIS | Alibaba |
| I2VGen-XL [100] | 3.1K | 11'23 | TV | IV | VV | | 30 | 8 | 200 | S | Web-Vid10M, LAION-400M | Alibaba |
| SEINE [11] | 1K | 11'23 | TV | | | | 2 | 8 | 16 | S | UCF101 | Shanghai AI |
| SVD [29] | 25.9K | 11'23 | TV | IV | | | 2 | 14 | 25 | S | UCF101, LVD-10M, MVImgNet, Google Scanned Objects | Stability AI |
| MAGVIT-v2 [97] | - | 10'23 | TV | IV | | | 2 | 8 | 17 | S | UCF-101, Kinetics-600, SSv2 | Google |
| AnimateDiff [23] | 11.4K | 07'23 | | IV | | | 2 | 16 | 32 | S | WebVid-10M, Civitai | Stanford Uni |
| VDM [28] | - | 04'22 | TV | IV | | | 2 | 8 | 16 | S | UCF101, BAIR Robot Pushing, Kinetics-600 | Google |