Generative Action Tell-Tales: Assessing Human Motion in Synthesized Videos

Anonymous CVPR submission

Paper ID

Abstract

Despite rapid advances in video generative models, robust metrics for evaluating visual and temporal correctness of complex human actions remain elusive. Critically, existing pure-vision encoders and Multimodal Large Language Models (MLLMs) are strongly appearance-biased, lack temporal understanding, and thus struggle to discern intricate motion dynamics and anatomical implausibilities in generated videos. We tackle this gap by introducing a novel evaluation metric derived from a learned latent space of real-world human actions. Through extensive experiments, we show that our metric achieves substantial improvement of more than **68**% *compared to existing state-of-the-art methods on our benchmark.*

1. Introduction

How do humans learn the right way to perform an action, like *walking* or *making a toast*? Since infancy, we learn by implicitly observing and explicitly being taught by others, grasping the flow of events and the physical laws governing human motions and actions [2]. This intuitive understanding allows humans to effortlessly recognize motion inconsistencies even in today's highly photorealistic generated videos [31, 36, 41]. Current metrics for judging generated videos, such as pixel-level similarity [19, 43], perceptual quality [39, 47], or text-to-video prompt alignment [17, 26], as well as recent approaches that use MLLMs as judges [6, 15], do not accurately capture the complex motion physics of human actions [11, 37]. A core contribution of our work is to bridge this gap by building a robust assessment tool that moves beyond superficial statistics and bakes in the critical awareness of physical and anatomical-consistency of human motion.

Our key idea is to learn a latent manifold of natural human motion, built from semantic features that measure the consistency of human anatomy, motion physics, and visual appearance in a video. To evaluate a new video, we project its embeddings into this manifold and systematically measure the deviations and discern telltale signs of poor action correctness and coherence. While several benchmarks exist [26, 48] for evaluating video generative models, they fail to adequately probe for the fine-grained temporal correctness and coherence of human actions. We identify this as a critical limitation and create an open-source benchmark we call the "**T**elltale **A**ction **G**eneration Bench" (**TAG-Bench**), evaluating videos on: (a) whether the generated video captures the intended action, and (b) the temporal smoothness and anatomical plausibility of the perceived action.

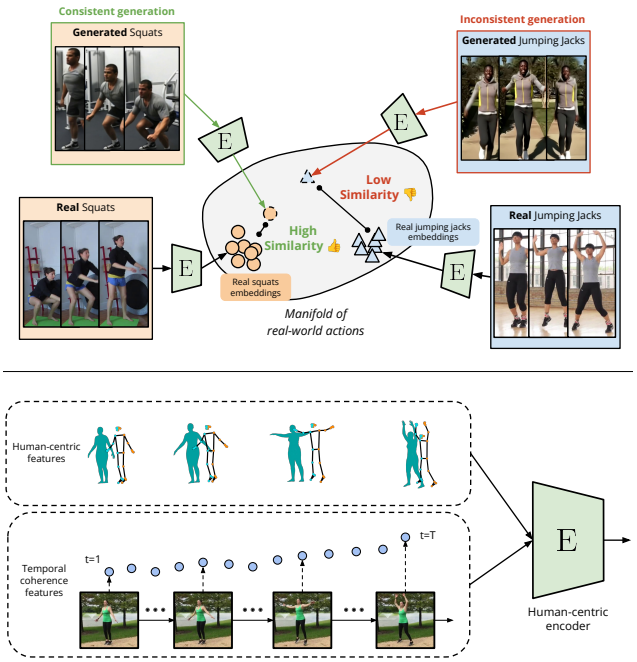

Figure 1. What are the **telltale signs of a generative action video**? We answer this by learning a robust manifold based on appearance and anatomical coherence exhibited by humans performing actions across several real-world videos. This manifold serves as anchors against which we project the features of a generated video in question and assess its realism.

2. Related Work

Video quality metrics. Metrics such as FVD [39] capture coarse spatiotemporal alignment but miss fine-grained semantic or physical accuracy. Frame-level measures including PSNR [19], SSIM [43], and LPIPS [47] assess visual

similarity to reference frames but ignore temporal coherence. CLIPScore [17] only captures single-frame semantics, lacking motion or physical plausibility. More recent MLLMs-based methods like VideoScore [15] and VideoPhy [6] predict human ratings or assess physical laws, but neither captures human motion or action correctness, underscoring the need for our proposed metric.

3. Approach

We posit the *"realism"* of human actions in generated videos as the distance between real and generated samples within a learned representation space, where real human actions cluster into a compact region of natural, physically plausible movements (Fig. 1).

3.1. Human-centric feature representations

We capture the complexity of human motion leveraging multiple human-centric features encompassing appearance, skeletal geometry, and motion dynamics.

(i) 3D features. We employ SMPL [27], a standard 3D body representation deformed by pose (θ), body shape (β), and global orientation (go). To infer these *features* from 2D video frames, we rely on human mesh recovery (HMR) [20] models. SMPL features are invariant to appearance and scene context, focusing specifically on body geometry, making them powerful ingredients to assess whether generated humans follow real-world physical dynamics. **(ii) 2D features.** SMPL is trained on real human data [27], constraining its parameters to anatomically plausible configurations, potentially overlooking anomalies such as elongated limbs common in generated videos. Thus, we also incorporate 2D joint keypoints [9] (kp_{2D}), which complement the 3D features by revealing anatomical distortions that SMPL might overlook. **(iii) Visual appearance features.** To complement the appearance-invariant 3D and 2D features, we extract visual appearance features (f_{vis}) using pre-trained image-based visual backbones (e.g., ViT [13]), capturing cues such as clothing, color, and action-relevant objects. **(iv) First-order temporal coherence.** We refer to the above features ($\mathbb{S} = \{\theta, \beta, go, kp_{2D}, f_{vis}\}$) as *static*, as they capture body state at a single frame. We additionally compute their first-order temporal derivatives, yielding *motion* features ($\mathbb{U} = \{m_\theta, m_\beta, m_{go}, m_{kp_{2D}}, m_{f_{vis}}\}$), making the manifold sensitive to artifacts such as sudden body shape changes, jitter, or implausible pose transitions.

3.2. Learning a manifold of real-world actions

Our goal is to learn a compact human-centric *latent representation* where physically plausible actions occupy compact regions, while anatomically distorted or temporally inconsistent actions lie farther apart. We train an encoder to distinguish physically plausible motion from implausible human actions, described next.

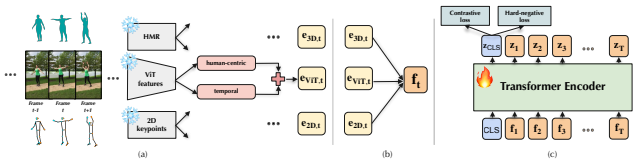

Figure 2. **Architectural overview of the encoder we train to learn the real-world action manifold.** We extract per-frame static human-centric and temporal motion features (Fig. (a)) (Sec. 3.1), and aggregate them, yielding one embedding for each frame (Fig. (b)) (Sec. 3.2.1). We prepend a [CLS] token to the per-frame tokens and pass as input to a 4-layer transformer encoder (Fig. (c)) (Sec. 3.2.1). Our aim is to encourage the encoder to group diverse videos pertaining to a given action closer together. We also ensure that temporally incoherent videos lie farther apart.

3.2.1. Encoder architecture and training

Constructing temporal windows: We represent each video as a sequence of fixed-length *temporal windows* of T consecutive frames, extracting human-centric features from Sec. 3.1 for each frame. The model operates in three stages: (i) encode each input feature independently, (ii) fuse per-input representations per frame, and (iii) temporally aggregate frame-level representations over the window.

(i) Per-input encoding: We define two pathways for *static* ($\mathbb{S}$) and *motion* ($\mathbb{U}$) features, processed via separate 1D temporal convolution blocks ϕ^k_{static} and ϕ^k_{motion}. Static and motion encodings are combined via element-wise addition:

$$\mathbf{e}_{k,t} = \phi^k_{\text{static}}(s_{k,t}) + \phi^k_{\text{motion}}(u_{k,t}) \tag{1}$$

(ii) Per-frame feature fusion: We aggregate encoded features $\mathbf{e}_{k,t}$ into a single frame-level representation ($\mathbf{f}_t$) via learned a learned attention mechanism:

$$\mathbf{f}_t = \sum_k \alpha_{k,t}\, \mathbf{e}_{k,t}, \quad \alpha_{k,t} = \text{softmax}_k\left(\frac{\mathbf{q}^\top \mathbf{W}_a \mathbf{e}_{k,t}}{\sqrt{d}}\right) \tag{2}$$

(iii) Temporal aggregation: We prepend a learnable [CLS] token [8] and process the sequence with a Transformer encoder (Fig. 2), yielding a compact window-level embedding $\mathbf{z}_{\text{CLS}}$ and per-frame embeddings $\{\mathbf{z}_t\}_{t=1}^T$.

3.2.2. Training objective

We train with a multi-loss objective combining two goals: **(i) Action semantics:** A supervised contrastive loss ($\mathcal{L}_{\text{supcon}}$) [5] clusters window-level embeddings ($\mathbf{z}_{\text{CLS}}$) from the same action class while pushing apart those from different classes. **(ii) Temporal coherence:** We simulate temporally distorted variants of real videos by shuffling frames, repeating the first frame, or reversing frame order. A hard-negative loss ($\mathcal{L}_{\text{hard-negative}}$) pushes embeddings of distorted windows away from their coherent counterparts, teaching the model to recognize plausible motion dynamics. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{\text{supcon}} + \lambda \mathcal{L}_{\text{hard-negative}} \tag{3}$$

3.3. Quantitative Metrics

From this learned embedding space, we derive two quantitative measures based on key observations: real videos of the same action form compact **action consistent** clusters, and frame-level embeddings of a real video evolve smoothly, measuring **temporal coherence**.

To measure **action consistency (S_{cons})**, we compute an action-specific centroid $\mathbf{c}_k$ by averaging window-level embeddings across all real videos of class k. For a generated video, we similarly average its [CLS] embeddings across temporal windows to obtain $\mathbf{z}^{\text{gen}}_{\text{video}}$. Lower distance $S_{\text{cons}} = \left\|\mathbf{z}^{\text{gen}}_{\text{video}} - \mathbf{c}_k\right\|_2$ indicates closer alignment with real action distributions.

To evaluate **temporal coherence (S_{temp})**, we measure the smoothness of frame trajectories within the learned embedding space:

$$S_{\text{temp}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2. \tag{4}$$

Lower scores correspond to gradual, physically consistent transitions, while higher scores indicate abrupt or implausible temporal changes.

4. Telltale Action Generation (TAG)-Bench

Dataset construction: We select 10 action classes from UCF-101 [35] featuring a single visible person depicting diverse whole-body movements: *BodyWeightSquats, HulaHoop, JumpingJack, PullUps, PushUps, Shotput, SoccerJuggling, TennisSwing, ThrowDiscus,* and *WallPushups*. For each action, we sample 6 videos and extract their first frame as input to image-to-video (I2V) models – Wan2.1 [41], Wan2.2 [41], Hunyuan [21], Opensora [49], and Runway Gen-4 [36] – with the prompt "A person is doing {action}," yielding 300 generated videos. We focus on I2V to isolate differences in motion generation without confounding factors like scene layout or appearance.

Video annotation setup: We recruit 246 participants through Amazon Mechanical Turk [4] to rate videos on a 1–10 scale along two axes: (1) ***Action Consistency***, how accurately the video depicts the intended action, and (2) ***Temporal Coherence***, how physically plausible and temporally smooth the motion appears. After subject rejection, inter-rater correlation reached 0.716 for *Action Consistency* and 0.710 for *Temporal Coherence*.

Method	Corr. with Action Consistency ↑	Corr. with Temporal Coherence ↑
Random	-0.07	-0.11
Feature-based automatic metrics		
PIQUE [40]	-0.19	-0.13
BRISQUE [30]	-0.04	0.01
CLIP-sim [34]	0.03	0.16
DINO-sim [10]	0.08	0.21
SSIM-sim [43]	-0.08	-0.04
MSE-dyn [42]	-0.15	-0.08
SSIM-dyn [42]	-0.06	-0.03
CLIP-Score [17]	0.08	0.00
X-CLIP-Score [29]	0.00	-0.07
TRAJAN [3]	-0.12	-0.12
VideoMAE(UCF101)-classification	0.18	0.17
VBench-2.0 [48] (Human Anatomy)	-0.40	0.02
VBench-2.0 [48] (Human Identity)	0.06	0.02
VBench-2.0 [48] (Human Clothes)	0.12	0.11
MLLM-based fine-tuned metrics		
🔓 VideoScore [15] (Visual Quality)	-0.12	-0.06
🔓 VideoScore [15] (Temporal Consistency)	-0.09	-0.04
🔓 VideoScore [15] (Dynamic Degree)	-0.19	-0.16
🔓 VideoScore [15] (T2V Alignment)	-0.07	-0.04
🔓 VideoScore [15] (Factual Consistency)	-0.14	-0.08
🔓 VideoScore2 [16] (Visual Quality)	0.14	0.16
🔓 VideoScore2 [16] (T2V Alignment)	0.17	0.09
🔓 VideoScore2 [16] (Physical Consistency)	0.18	0.17
🔓 VideoPhy-2 [7] (Semantic Adherence)	0.19	0.16
🔓 VideoPhy-2 [7] (Physical Commonsense)	0.28	0.37
MLLM Prompting		
🔓 LLaVA-1.5-7B [24]	-0.17	-0.14
🔓 LLaVA-v1.6-mistral-7b-hf [25]	-0.10	0.18
🔓 Idefics2-8B [23]	-0.05	-0.06
🔓 Qwen3-VL-8B-Instruct [45]	0.34	0.28
🔒 Gemini-2.5-Flash [12]	0.40	0.25
🔒 Gemini-2.5-Pro [12]	0.31	0.26
🔒 GPT-4o [1]	0.34	0.31
🔒 GPT-5 [33]	<u>0.45</u>	<u>0.38</u>
Ours		
Action Consistency S_{cons} (Ours)	**0.61**	0.45
Temporal Coherence S_{temp} (Ours)	0.53	**0.64**
Δ over best baseline	+ 0.16	+ 0.26
Relative improvement (%) over best baseline	+ 35.6%	+ 68.4%
Inter-rater agreement		
Human vs Human	0.72	0.71

Table 1. **Correlation (Spearman's ρ) between model predictions and human scores for *Action Consistency* and *Temporal Coherence*.** (Higher is better). 'VideoMAE(UCF101)-classification' uses the confidence score [38] as the predicted scores. 🔓 denotes open-source models, while 🔒 denotes closed-source models. We observe that the proposed S_{cons} outperforms all methods for *Action Consistency*, and S_{temp} for *Temporal Coherence*. The next best performing metric is <u>underlined</u>. Details in Appendix.

5. Experiments

5.1. Implementation details

Features. We extract SMPL parameters using TokenHMR [14] and 2D keypoints using DWPose [46]. Visual appearance features are obtained from the frozen ViT-H/16 backbone of TokenHMR. Motion features are computed as frame-to-frame differences: relative rotations for pose and global orientation, and ℓ_2 distances for appearance, shape, and keypoint features. All features are flattened and normalized prior to being passed to the model.

Model details. Each feature is processed by a 1D temporal convolution block with three layers (kernel size 5, dilation $\{1, 2, 4\}$) and residual connections, outputting a 256-D embedding (Fig. 2(a)). These are fused via attention into a single 256-D per-frame representation (Fig. 2(b)). A learnable [CLS] token with sinusoidal positional embeddings is prepended and processed by a 4-layer Transformer encoder (8 heads), yielding ℓ_2-normalized $\mathbf{z}_{\text{CLS}}$ and $\{\mathbf{z}_t\}$ (Fig. 2(c)).

Training. We train on real videos of the same 10 UCF-101 action categories as TAG-Bench, using the default setting of 320×240 resolution videos at 25 FPS divided into windows of T=32 frames and overlap of 8 frames between windows. Videos used for TAG-Bench generation (Sec. 4) or containing multiple people are excluded, with the remainder split 80/20 for training/validation. We train for 90 epochs using AdamW [28] (lr 3×10^{-4}, weight decay 1×10^{-4}, batch size 256, λ=10 (Eqn. 3)) on a single A100 GPU.

Evaluation. We report Spearman's rank correlation (ρ) between metrics and human ratings [15].

5.2. Comparison to automatic metrics and MLLMs

We evaluate a diverse set of automatic video quality metrics on TAG-Bench, assessing alignment with human ratings of action correctness and motion quality (Table 1).

Prior methods. Frame-level metrics like CLIP-Score [17] fail to capture motion dynamics, correlating poorly with human ratings (< 0.22), while TRAJAN [3] and VBench-2.0 [48] metrics also underperform (< 0.11), indicating that scene coherence and per-frame anatomical evaluation do not imply accurate action. MLLM-based evaluators like VideoScore [15, 16] and VideoPhy-2 [7] target general criteria such as visual fidelity or object-level physics rather than fine-grained human motion, achieving only 0.28 on *Action Consistency* and 0.37 on *Temporal Coherence*.

Prompting MLLMs. Directly using raw video inputs proved unreliable, as MLLMs often fail to capture fine-grained temporal details [44]. For instance, Gemini-2.5-Pro [12] shows low correlations of 0.25 for *Action Consistency* and 0.22 for *Temporal Coherence* when passed as videos. To mitigate this, we uniformly sample 40 frames from each video arranged into 4×10 grid panels, preserving both temporal progression and spatial structure. This improves alignment: Gemini-2.5-Pro's correlations rise to 0.31 on *Action Consistency* and 0.26 on *Temporal Coherence*. Among MLLMs, GPT-5 achieves the highest alignment of 0.45 for *Action Consistency* and 0.38 for *Temporal Coherence* (Table 1).

Proposed metrics. Our metrics (Sec. 3.3) show stronger alignment with human perception (Table 1): *Action Consistency* (S_{cons}) achieves 0.61 and *Temporal Coherence* (S_{temp}) achieves 0.64, outperforming GPT-5 by $+\mathbf{35.6}\%$ and $+\mathbf{68.4}\%$ in relative gains respectively.

5.3. Performance on external benchmarks

We evaluate on the Human Anatomy subset of VBench-2.0 [48], which compares videos from four *text-to-video* models (Sora-480p [32], Kling [22], Hunyuan [21], and CogVideo [18]) via pairwise human preference comparisons. We evaluate only videos with a single visible person and compute *win ratios* [48], ranking models by the fraction of pairwise comparisons they win. Since VBench-2.0 prompts do not correspond to our 10 training classes, we evaluate using only S_{temp}, as S_{cons} requires a corresponding action centroid. We find that S_{temp} produces the same ranking of models as human raters, despite being evaluated on text-to-video models with unseen action prompts, demonstrating that the learned embedding generalizes beyond training.

5.4. Analysis

Effect of loss terms. Removing $\mathcal{L}_{\text{supcon}}$ causes a steep drop in *Action Consistency* (0.61 → 0.26) and *Temporal Coherence* (0.64 → 0.38) (Table 2), validating that action semantics are essential for learning plausible motion. Adding $\mathcal{L}_{\text{hard-negative}}$ further improves both scores (*Action Consistency* : 0.54 → 0.61, *Temporal Coherence* : 0.57 → 0.64), confirming that both losses are essential.

$\mathcal{L}_{\text{supcon}}$	$\mathcal{L}_{\text{hard-neg}}$	Action Consistency	Temporal Coherence
✓	✓	**0.61**	**0.64**
✗	✓	0.26	0.38
✓	✗	0.54	0.57

Table 2. **Effect of the loss terms** $\mathcal{L}_{\text{supcon}}$ and $\mathcal{L}_{\text{hard-neg}}$. Both terms jointly improve action consistency and temporal coherence; removing either hurts.

Pose	Body shape	Global orientation	Keypoints	Visual features	Motion	Action Consistency	Temporal Coherence
✓	✓	✓	✓	✓	✓	**0.61**	**0.64**
✗	✓	✓	✓	✓	✓	0.56	0.57
✓	✗	✓	✓	✓	✓	0.54	0.57
✓	✓	✗	✓	✓	✓	0.57	0.57
✓	✓	✓	✗	✓	✓	0.61	0.57
✓	✓	✓	✓	✗	✓	0.56	0.59
✓	✓	✓	✓	✓	✗	0.46	0.50

Table 3. **Effect of each input feature.** We report Spearman's correlation (ρ) with human scores after zeroing each input feature independently. Models are retrained from scratch for each setting. "Motion" denotes temporal derivatives of all inputs (Sec. 3.1). Removing motion causes the largest degradation.

Ablation on input features. Zeroing out (while retaining full dimensionality of the input) any single feature decreases performance (Table 3); masking 3D pose drops *Action Consistency* from 0.61 → 0.56, while masking all motion features causes the largest drop (*Action Consistency* : 0.61 → 0.46), highlighting the necessity of our multi-feature representation.

Extending TAG-Bench to more action classes. We extend TAG-Bench from 10 to 23 classes by incorporating 13 additional UCF-101 [35] actions (*Bowling, Clean&Jerk, GolfSwing, HammerThrow, Hammering, HandStandPushup, JugglingBalls, JumpRope, Lunges, PlayingGuitar, RockClimbingIndoor, RopeClimbing,* and *Surfing*), following the same evaluation protocol as Sec. 4. Our model continues to outperform, achieving *Action Consistency*: 0.59 and *Temporal Coherence*: 0.56, compared to GPT-5's *Action Consistency*: 0.46 and *Temporal Coherence*: 0.39.

6. Discussion and Future Work

We present a framework for evaluating human actions in generated videos by decomposing the task into two aspects: action consistency against real videos and the temporal smoothness of human motion. We demonstrate strong alignment with human perceptual judgment and generalizability across diverse generation models. While our main study is conducted on 10 action classes, it offers a proof-of-concept for the utility of the proposed features. Sec. 5.4 shows the extensibility of the framework to more actions. Future work will extend this framework to more actions, longer-form videos, and in-depth exploration of integrating human-physics-based features with modern MLLMs.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Alison Gopnik, Patricia K. Kuhl, and Andrew N. Meltzoff. The scientist in the crib: Minds, brains, and how children learn, 1999. 1

[3] Kelsey Allen, Carl Doersch, Guangyao Zhou, Mohammed Suhail, Danny Driess, Ignacio Rocco, Yulia Rubanova, Thomas Kipf, Mehdi SM Sajjadi, Kevin Murphy, et al. Direct motion models for assessing generated videos. *arXiv preprint arXiv:2505.00209*, 2025. 3, 4

[4] Amazon Web Services, Inc. Amazon mechanical turk. https://www.mturk.com/. Accessed: November 2025. 3

[5] Chaitanya Animesh and Manmohan Chandraker. Tuned contrastive learning. *arXiv preprint arXiv:2305.10675*, 2023. 2

[6] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 1, 2

[7] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 3, 4

[8] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *International Conference on Machine Learning*, 2021. 2

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[11] Joseph Cho, Fachrina Dewi Puspitasari, Sheng Zheng, Jingyao Zheng, Lik-Hang Lee, Tae-Ho Kim, Choong Seon Hong, and Chaoning Zhang. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024. 1

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 3, 4

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[15] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. 1, 2, 3, 4

[16] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025. 3, 4

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. 1, 2, 3, 4

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 4

[19] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *Proceedings of the 20th IEEE International Conference on Pattern Recognition*, 2010. 1

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[21] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3, 4

[22] Kuaishou Technology. Kling: High-fidelity and temporally consistent text-to-video generation. *Technical Report*, 2024. https://kling.kuaishou.com. 4

[23] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 2024. 3

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023. 3

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge for large multimodal models (llava v1.6). https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024. Accessed: 2025-11-10. 3

[26] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[29] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval, 2022. 3

[30] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012. 3

[31] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025. 1

[32] OpenAI. Sora: A large-scale diffusion transformer for text-to-video generation. *Technical Report*, 2024. https://openai.com/research/sora. 4

[33] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. Accessed: 2025-11-10. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PmLR, 2021. 3

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 4

[36] Runway Research Team. Runway gen-4: Advancing realistic text-to-video generation. *Technical Report*, 2024. https://research.runwayml.com/gen4. 1, 3

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1

[38] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in Neural Information Processing Systems*, 2022. 3

[39] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[40] Narasimhan Venkatanath, D. Praneeth, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception-based features. In *Proceedings of the 2015 Twenty First National Conference on Communications*, 2015. 3

[41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3

[42] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 3

[43] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 1, 3

[44] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025. 4

[45] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[46] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1

[48] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 1, 3, 4

[49] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3