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

Xavier Thomas[1], Youngsun Lim[1], Ananya Srinivasan[2], Audrey Zheng[3], and Deepti Ghadiyaram[1,4]

[1] Boston University
[2] Belmont High School
[3] Canyon Crest Academy
[4] Runway

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. Our method first captures the nuances, constraints, and temporal smoothness of real-world motion by fusing appearance-agnostic human skeletal geometry features with appearance-based features. We posit that this combined feature space provides a robust representation of action plausibility. Given a generated video, our metric quantifies its action quality by measuring the distance between its underlying representations and this learned real-world action distribution. For rigorous validation, we develop a new multi-faceted benchmark specifically designed to probe temporally challenging aspects of human action fidelity. 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, performs competitively on established external benchmarks, and has a stronger correlation with human perception. Our in-depth analysis reveals critical limitations in current video generative models and establishes a new standard for advanced research in video generation. Code is available at https://xthomasbu.github.io/video-gen-evals/

Keywords: Video Generation · Evaluation Metrics · Human Action Consistency · Temporal Consistency

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

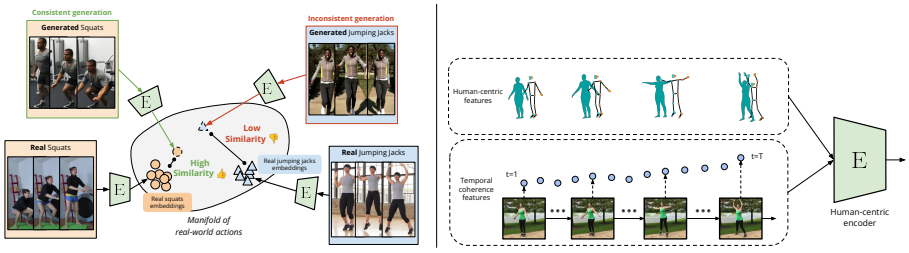


Fig. 1: What are the **telltale signs of a generative action video**? We answer this by learning a robust manifold based on appearance and anatomical coherence exhibited by humans performing actions across several real-world videos. This manifold serves as anchors against which we project the features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

effortlessly recognize motion inconsistencies even in today's highly photorealistic generated videos [34, 42, 47].

This raises a critical question: *Can we formalize our intuitive perception and design a framework to systematically evaluate the accuracy of human actions in generated videos?* This is a challenging task, as it requires solving a twofold challenge. First, the baseline problem of recognizing action correctness is ill-posed, even in real videos. Actions can be atomic [16] (e.g., "walking") or procedural (e.g., "making a toast") [7], making automatic recognition inherently challenging. Second, for generated videos, this problem is magnified as we must move beyond detecting the mere **presence of an action** to also evaluating the **temporal coherence** of the body movements over time.

Current metrics for judging generated videos, such as pixel-level similarity [23,48], perceptual quality [45,54], or text-to-video prompt alignment [20,30], as well as recent approaches that use MLLMs as judges [5,19], do not accurately capture the complex motion physics of human actions [12, 43]. We posit that this is because the underlying representations fundamentally lack awareness of subtle anatomical violations or temporal incoherence. 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. This manifold is built from semantic features that measure the consistency of human anatomy, motion physics, and visual appearance in a video. As illustrated in Fig. 1, we learn this manifold from a large volume of diverse videos capturing a wide range of body geometries and performance styles (e.g., a tall woman walking vs. an older male with a walking stick), thereby encapsulating the crucial cues of natural action. To evaluate a new video, we project its embeddings into this manifold and systematically measure the deviations and discern telltale signs of poor action quality, correctness, and coherence.

While several benchmarks exist [30,56], 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-v0**). We generate hundreds

of videos from state-of-the-art open- and closed-source models and conduct a large-scale subjective study. We evaluate the videos on two key criteria: (a) whether the generated video captures the intended action, and (b) the temporal smoothness and anatomical plausibility of the perceived action.

Our extensive experiments on **TAG-Bench-v0** and other benchmarks [56] yield several key findings. First, we highlight a critical gap in current evaluation: we show that all state-of-the-art models and MLLMs struggle to correctly evaluate action correctness and temporal coherence. This demonstrates the sheer difficulty of the task. We further find that certain actions are challenging for all generative models, revealing broader limitations in current video synthesis capabilities. Second, we demonstrate that, unlike existing methods, our manifold's scores strongly align with human opinion, across both image-to-video (I2V) [55] and text-to-video (T2V) [9] generation settings, and across multiple models. In summary, our contributions are:

- We design a learned latent space that encodes human body geometry and temporal coherence to evaluate action quality in generated videos.
- We design **TAG-Bench-v0**, a benchmark with human ratings focused on the correctness and temporal coherence of human actions in generated videos.
- We propose two metrics that align closely with human perception. They measure the consistency and the temporal plausibility of actions.
- We provide an in-depth analysis of the proposed latent action manifold, and validate the design choices that led to its robust performance.

2 Related Work

Video generation models. Video generation [22, 27, 35, 42, 47] has rapidly advanced, producing increasingly photorealistic and temporally coherent content [21, 51]. Beyond visual fidelity, recent work [49] frames these models as emerging *world models* capable of capturing complex real-world dynamics without explicit supervision. Despite such progress, current systems still often generate human motion that is kinematically implausible or semantically incorrect [12, 43]. This highlights the need for metrics that specifically diagnose and quantify these failures.

Distribution and reference based metrics. Metrics such as FVD [45] measure statistical similarity between real and generated videos in pretrained feature spaces, capturing coarse spatiotemporal alignment but missing fine-grained semantic or physical accuracy. Frame-level measures including PSNR [23], SSIM [48], and LPIPS [54] assess visual similarity to reference frames but ignore temporal coherence critical to video perception. Recent efforts like Physics-IQ [34] measure spatiotemporal realism using real-world references, but do not focus on capturing human-body distortions or action-level correctness.

Reference-free video metrics. In many scenarios, ground-truth videos are unavailable, motivating the need for reference-free evaluation. CLIPScore [20] measures frame-text similarity scores but only captures single-frame semantics, lacking motion or physical plausibility. More recent methods employ MLLMs for

richer video understanding, via zero-shot prompting or fine-tuning on human ratings. For instance, VideoScore [19] predicts fine-grained human ratings across dimensions such as visual quality and temporal consistency, while VideoPhy [5] assesses whether generated videos follow physical laws like gravity or buoyancy. However, neither captures human-body distortions or action correctness, underscoring the need for our proposed metric.

Benchmarking video models. With advances in generative video, several benchmarks now evaluate videos across multiple dimensions. EvalCrafter [30] provides a large-scale framework for assessing visual quality, motion quality, temporal consistency, and text–video alignment. VBench2.0 [56] extends this with finer-grained anatomy-related criteria. However, none explicitly assess whether human actions are executed plausibly over time. Our proposed TAG-Bench-v0 addresses this gap through targeted metrics and a curated evaluation set designed to quantify human action realism and physically plausible motion.

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. In this space, real human actions cluster into a compact region of natural, physically plausible movements. We refer to this as the action manifold (Fig. 1). Capturing this notion of realism requires accurately modeling the *temporal intrinsics* of the human body, human-object interactions, and the sequence of atomic actions involved in performing a given action.

This section is structured as follows: we detail the variety of human-centric features in Sec. 3.1 that serve as building blocks of the learned action mainfold, which we describe in Sec. 3.2. Next, we define the distance metric we learn to assess realism in generated videos in Sec. 3.3.

3.1 Human-centric feature representations

We capture the complexity of human motion leveraging human-centric features that measure appearance, skeletal geometry, and motion dynamics detailed next.

3D features. We employ Skinned Multi-Person Linear (SMPL) [31], a standard 3D representation of the human body. Briefly, SMPL consists of a "rest pose", a 3D mesh representation of an average human body, that is deformed based on three parameters: pose (θ), body shape (β), and global orientation (go). θ represents the 3D rotations of skeletal joints [31], capturing how the human body moves during a particular action. β captures shape-specific characteristics [31], such as overall build (e.g., tall vs. short) and limb proportions. go describes the global body rotation computed from the pelvis joint [24], capturing how the orientation of the person's entire body changes during an action (e.g., turning sideways). To infer these SMPL features from 2D video frames, we rely on human mesh recovery (HMR) [25] models.

Motivation to use SMPL. We believe that the detailed 3D features are crucial to capture the complex kinematics of a human body in action. SMPL features are invariant to appearance and scene context [26,31], and focus specifically on the geometry of the human body. They serve as powerful ingredients to assess if generated humans follow the same physical dynamics as real-world humans.

2D features. While 3D representations capture detailed joint rotations and body shapes, SMPL is trained solely on real human data [31], constraining its parameters to remain within anatomically plausible body configurations [38]. This may overlook anomalies such as elongated limbs or implausible joint configurations common in *generated* videos. Thus, we also incorporate 2D joint keypoints [10] (kp_{2D}), which are void of such strong priors. These 2D cues complement the 3D features by revealing anatomical distortions that the SMPL representation might overlook.

Visual appearance features. Although 3D and 2D features capture body structure, they are inherently and intentionally appearance-invariant [10,31,39]. To complement them, we extract visual appearance features (f_{vis}) using pretrained image-based backbones (e.g., ViT [14]). These features capture cues such as clothing, color, and action-relevant objects, all of which influence how an action is visually perceived.

First-order temporal coherence. We refer to the human-centric features described so far ($\mathbb{S} = \{\theta, \beta, go, kp_{2D}, f_{vis}\}$) as *static* features, as they capture the body's state at a particular point in time (i.e., a frame). However, human actions inherently involve body dynamics that temporally evolve in a coherent manner. Consider a generated video of a person performing bar pull-ups where the person's arms become unrealistically long over time. We believe that this unnatural body morphing will strongly emerge when we compute the temporal derivative of the body shape feature and will serve as a critical signal. Motivated by this, we compute the first-order temporal derivatives of each static feature, yielding corresponding *motion* features: ($\mathbb{U} = \{m_\theta, m_\beta, m_{go}, m_{kp_{2D}}, m_{f_{vis}}\}$). These features make the action manifold sensitive to generative artifacts such as sudden body shape changes, jitter, or implausible pose transitions.

By combining 3D and 2D skeletal and appearance features and their corresponding temporal derivatives, we obtain a human-centric representation that is anatomically grounded and captures the natural evolution of human motion. Yet, we acknowledge that HMR and pose features may be unreliable under occlusion.

3.2 Learning a manifold of real-world actions

Our goal is to learn a compact human-centric *latent representation* that captures the nuances of real-world actions. In this space, physically plausible actions occupy compact regions, while anatomically distorted or temporally inconsistent

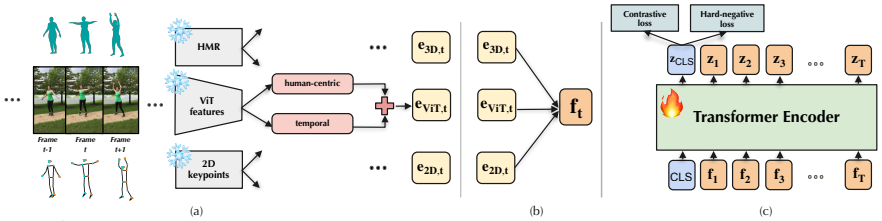


Fig. 2: Architectural overview of the encoder we train to learn the real-world action manifold. We extract per-frame static human-centric and temporal motion features (Fig. (a)) (Sec. 3.1), and aggregate them, yielding one embedding for each frame (Fig. (b)) (Sec. 3.2). We prepend a [CLS] token to the per-frame tokens and pass as input to a 4-layer transformer encoder (Fig. (c)) (Sec. 3.2). 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.

actions lie farther apart. To this end, we train an encoder to distinguish physically plausible motion from implausible human actions, as described next.

Encoder architecture and training. We represent each video as a sequence of fixed-length *temporal windows*, each containing T consecutive frames. This design allows the model to capture local, fine-grained motion of an action (e.g., a stride of a person running). For each frame $t \in \{1, \ldots, T\}$ in a window, we extract human-centric features from Sec. 3.1, which serve as input to our model (Fig. 2(a)). The proposed model operates in three stages: (i) encode each input feature independently, (ii) fuse resulting per-input representations for each frame, and (iii) temporally aggregate these frame-level representations over the temporal window. We describe each stage in detail next.

(i) Per-input encoding: We take inspiration from [40] and define two pathways, one for *static* ($\mathbb{S}$) and another for *motion* ($\mathbb{U}$) features (Sec. 3.1). To this end, let ϕ denote a 1D temporal convolution block that aggregates features within a short temporal context. Each static and motion input feature ($s_{k,t} \in \mathbb{S}$ and $u_{k,t} \in \mathbb{U}$ respectively, where k represents the distinct feature sources (i.e., pose, shape, keypoints, etc) in Sec. 3.1) is processed using separate temporal convolution blocks, ϕ^k_{static} and ϕ^k_{motion}. These per-input temporal convolutions help capture the local temporal evolution of each input feature (e.g., θ).

Next, we combine the static and motion encodings for each frame t via element-wise addition, similar to the additive fusion used in GENMO [29]. Thus, the motion pathway provides residual information to the static pathway:

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

where $u_{k,t}$ is the temporal derivative of the corresponding static feature $s_{k,t}$, and $\mathbf{e}_{k,t} \in \mathbb{R}^d$, where d is the dimension of the encoded input feature.

(ii) Per-frame feature fusion: We next aggregate the encoded input features $\mathbf{e}_{k,t}$ to obtain a single frame-level representation using a learned attention mechanism. Specifically, for each frame t and input k, the model assigns a scalar weight $\alpha_{k,t}$, capturing its relative importance for that frame. The fused representation

$\mathbf{f}_t$ is then computed as a weighted sum of all features:

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

The attention weights $\alpha_{k,t}$ are computed via a scaled dot-product attention [46] between a learnable query vector $\mathbf{q} \in \mathbb{R}^d$ and a linear projection of each input feature using $\mathbf{W}_a \in \mathbb{R}^{d \times d}$. Softmax operation ensures that the weights are positive and sum to 1. This attention mechanism is learned jointly with the model.
(iii) Temporal aggregation: To aggregate all the fused frame representations $\{\mathbf{f}_t\}_{t=1}^T$ for a given temporal window, we prepend a learnable token denoted as [CLS] [8], and process the sequence with a Transformer encoder, which models long-range temporal dynamics (Fig. 2(c)). This results in a compact embedding $\mathbf{z}_{\text{CLS}}$ that captures the essential information over a temporal window. In addition, we extract the sequence of frame-level output embeddings $\{\mathbf{z}_t\}_{t=1}^T$ (Fig. 2(c)) to capture a finer-grained per-frame representation.

Training objective. Our goal is to learn an effective latent space of real-world human actions. We achieve this via two complementary losses:
(i) Learn action semantics: We use a supervised contrastive loss ($\mathcal{L}_{\text{supcon}}$) [4] which encourages window-level embeddings ($\mathbf{z}_{\text{CLS}}$) from the same action class to cluster together in the latent space, while pushing embeddings from different action classes apart. This results in representations that are discriminative across actions, facilitating a robust understanding of what action is being performed.
(ii) Enforcing temporal coherence: In addition to distinguishing between actions, we also want our learned manifold to be temporally sensitive. For instance, a generated "jumping jacks" video may contain correct poses yet appear unrealistic if the person remains frozen mid-motion intermittently. To enforce this property in a self-supervised manner, we simulate temporally distorted variants of real videos by: (i) shuffling frames (breaking motion continuity), (ii) repeating the first frame across the window (simulating static frames), and (iii) reversing frame order (disrupting causal progression). We then introduce an additional loss ($\mathcal{L}_{\text{hard-negative}}$) that penalizes temporal inconsistency by pushing embeddings of these distorted windows away from their temporally coherent counterparts, teaching the model to recognize plausible motion dynamics. Crucially, this objective goes beyond standard semantic negatives (i.e., different actions) and explicitly teaches the model to be sensitive to temporal structure.

The overall training loss is a weighted sum between the two objectives:

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

where λ is a scalar weighting coefficient that balances the two loss terms. This combined objective encourages the encoder to capture both action semantics (*what* action) and temporal coherence (*how* the action is performed).

3.3 Quantitative Metrics

From this learned embedding space, we derive quantitative measures to assess how closely a generated video aligns with the manifold of real human actions based on two key observations:

- Real videos of the same action (e.g., "jumping jacks") form compact, **action consistent** clusters (Fig. 1).
- Frame-level embeddings of a real video evolve smoothly over time, measuring **temporal coherence** (Fig. 1).

Given a presumably poorly generated video, it should violate these two properties. To measure **action consistency (S_{cons})**, we compute an action-specific centroid $\mathbf{c}_k$ by averaging the temporal window-level embeddings across all real videos of a given class k. For a generated video of the same action, we similarly average its window-level [CLS] embeddings to obtain $\mathbf{z}^{\text{gen}}_{\text{video}}$. A generated video is more action consistent if the distance between these two representations ($S_{\text{cons}} = \|\mathbf{z}^{\text{gen}}_{\text{video}} - \mathbf{c}_k\|_2$) is small. Lower distance 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. Given per-frame embeddings ($\mathbf{z}_t$) of a window, we define temporal coherence as

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

The final score for a video is obtained by averaging temporal coherence scores across all windows. Lower scores correspond to gradual, physically consistent transitions, while higher scores indicate abrupt or implausible temporal changes. While action consistency measures if a generated action is semantically correct, temporal coherence measures how temporally smooth the generation is. In Sec. 5.2, we show that these metrics strongly correlate with human perception and serve as reliable indicators for evaluating generative video models.

4 Telltale Action Generation (TAG)-Bench

As mentioned earlier, existing benchmarks [18, 56] do not probe for the finer-grained temporal correctness and coherence of human actions. To bridge this limitation, we build an open-source benchmark, "**T**elltale **A**ction **G**eneration (TAG)-Bench-v0". TAG-Bench-v0 comprises videos generated from several open- and closed-source I2V generation models associated with rich human opinion scores. We describe the data curation and human annotation process below.

Dataset construction: We select 10 action classes from UCF-101 [41] that (1) feature a single visible person, (2) depict the full human body, and (3) involve diverse whole-body movement dynamics: *BodyWeightSquats, HulaHoop, JumpingJack, PullUps, PushUps, Shotput, SoccerJuggling, TennisSwing, ThrowDiscus,* and *WallPushups*. By covering both localized (e.g., push-ups) and complex

Method	Corr. with Action Consistency ↑	Corr. with Temporal Coherence ↑
Random	-0.07	-0.11
Feature-based automatic metrics (Top-3)		
VideoMAE(UCF101)-classification	0.18	0.17
DINO-sim [11]	0.08	0.21
CLIP-sim [39]	0.03	0.16
MLLM-based fine-tuned metrics (Top-3)		
🔓 VideoPhy-2 [6] (Physical Commonsense)	0.28	0.37
🔓 VideoPhy-2 [6] (Semantic Adherence)	0.19	0.16
🔓 VideoScore2 [18] (Physical Consistency)	0.18	0.17
MLLM Prompting (Top-3)		
🔒 GPT-5 [36]	<u>0.45</u>	<u>0.38</u>
🔒 Gemini-2.5-Pro [13]	0.31	0.26
🔒 GPT-4o [1]	0.34	0.31
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 [44] 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>. We report the top-3 performing baselines for each category; more results in suppl.

full-body dynamics (e.g., tennis swings), the selected classes allow for a meaningful evaluation of human motion synthesis quality. For each action, we sample 6 videos at random and extract their first frame. This frame serves as the input to I2V generation models – Wan2.1 [47], Wan2.2 [47], Hunyuan [27], Opensora [57], and Runway Gen-4 [42] – along with the prompt "A person is doing {action}." This yields 300 generated videos (more in Appendix). We focus on I2V to ensure all models begin from the same visual input (i.e., the initial frame), allowing us to isolate differences in motion generation capabilities without confounding factors like variations in scene layout or the person's appearance.

Video annotation setup: We recruit 246 participants through Amazon Mechanical Turk [3]. Each participant rated the videos on a 1–10 scale along two axes: (1) ***Action Consistency***, i.e., how accurately the generated video depicts the intended action mentioned in the prompt and (2) ***Temporal Coherence***, i.e., how physically plausible and temporally smooth the motion appears in the generated video. After subject rejection, the average inter-rater correlation reached 0.716 for *Action Consistency* and 0.710 for *Temporal Coherence*. Additional details in Appendix.

5 Experiments

We outline the model implementation and training setup in Sec. 5.1. Next, we evaluate our proposed metrics against existing video evaluation metrics on TAG-

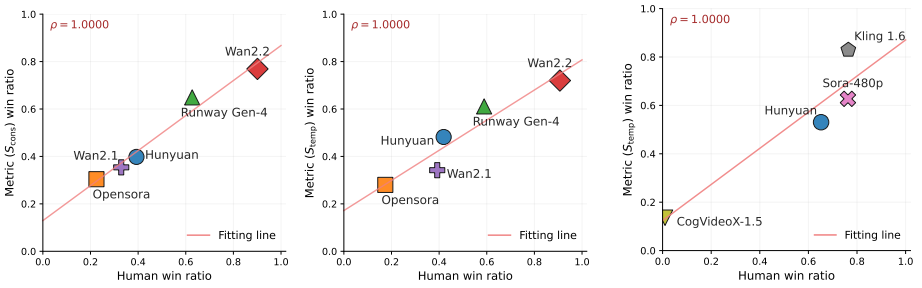
Bench-v0 (Sec 5.2) and external benchmarks (Sec 5.3). We then compare multiple open- and closed-source video generative models in Sec. 5.4, followed by an analysis on key design choices underpinning our approach in Sec. 5.5.

5.1 Implementation details

Features: We extract SMPL parameters using TokenHMR [15] (Sec. 3.1) and 2D keypoints using DWPose [53] (Sec. 3.1). TokenHMR employs Detectron2 [50] to obtain bounding boxes for the person; the cropped image is then used to infer SMPL parameters (Sec. 3.1). Visual appearance features (Sec. 3.1) are obtained from the frozen ViT-H/16 backbone of TokenHMR. Motion features (Sec. 3.1) are computed as frame-to-frame differences in feature values. Specifically, for pose and global orientation (rotation matrices), we compute relative rotations between adjacent frames to capture angular motion, while for appearance, body shape, and 2D keypoints, we use Euclidean (ℓ_2) distance. All features are flattened and normalized prior to being passed to the model.
Model details: Recall that each feature is first processed by a 1D temporal convolutional block (Sec. 3.2), which contains three sequential 1D convolution layers with kernel size 5 and respective dilation factors $\{1, 2, 4\}$. The dilated convolutions enable the model to efficiently capture both short- and mid-range temporal patterns. A residual connection is applied after each layer [17]. Each temporal convolutional block outputs a 256-dimensional embedding per input, which are then fused using an attention weighing mechanism to produce a single 256-D representation per frame. Next, a learnable 256-D [CLS] token is prepended to the sequence of 256-D frame representations, and sinusoidal positional embeddings are added. This sequence is processed by a 4-layer Transformer encoder with 8 attention heads, resulting in $\mathbf{z}_{\text{CLS}}$ and $\{\mathbf{z}_t\}_{t=1}^{T}$ which are then ℓ_2-normalized.
Training data: We train our encoder from scratch using the same 10 UCF101 action categories mentioned in the human evaluation (Sec. 4). We use videos at their native resolution (320×240) and frame rate (25 FPS), and divide into temporal windows of T=32 frames with an overlap of 8 frames between two windows. We exclude videos whose first frame was used as input for generating videos for TAG-Bench-v0 (Sec. 4). To extract only features of the human performing the action, we further discard videos containing more than one person. The remaining real videos are split into training (80%) and validation (20%).
Training details. We train the encoder for 90 epochs using AdamW [32] with a learning rate of 3×10^{-4}, weight decay of 1×10^{-4}, batch size 256, λ=10 in Eq. 3, and a cosine learning rate schedule, on a single A100 GPU.
Evaluation: Following prior work [18, 19], we report Spearman's rank correlation (ρ) between the metrics and human ratings.

5.2 Comparison to automatic metrics and MLLMs

In this section, we evaluate a diverse set of automatic video quality metrics on TAG-Bench-v0 (Sec. 4), including feature-based methods, fine-tuned MLLM



(a) *Action Consistency* (TAG-Bench-v0) **(b)** *Temporal Coherence* (TAG-Bench-v0) **(c)** *Temporal Coherence* (VBench-2.0)

Fig. 3: Model comparisons on TAG-Bench-v0 and VBench-2.0 Human Anatomy. We compare models pairwise for the same input prompt; for each pair, the model with the higher score (human or metric) is the winner. We then plot the win ratios (see Sec. 5.3) of human scores (x-axis) against win ratios from our metric (y-axis). Our metrics (S_{cons} and S_{temp}) observe the same ranking of models as humans on both benchmarks.

evaluators, and zero-shot approaches, assessing their alignment with human ratings of action correctness and motion quality (Table 1).

Feature-based metrics: Frame-level metrics like CLIP-sim [39] and DINO-dim [11] fail to capture motion dynamics, correlating poorly with human ratings (<0.22). We also note that the metrics from VBench-2.0 [56] designed to assess human anatomy and identity per frame also show very low correlations on our benchmark (< 0.11) in the Appendix.

MLLM-based fine-tuned evaluators: Recent methods like VideoScore2 [18, 19] and VideoPhy-2 [6] fine-tune MLLMs on human ratings to predict video quality. However, they target general criteria such as visual fidelity or text–video alignment, rather than fine-grained human motion. For instance, the *Physical Commonsense* score in VideoPhy-2 assesses scene- and object-level physics (e.g., gravity, collisions), but not human-specific dynamics such as joint coordination. As a result, its alignment with human ratings on our benchmark is modest, achieving only 0.28 on *Action Consistency* and 0.37 on *Temporal Coherence*.

Prompting MLLMs: We assess how well existing MLLMs align with human judgments. Directly using raw video inputs proved unreliable, as MLLMs often fail to capture fine-grained temporal details in human motion [52]. For instance, Gemini-2.5-Pro [13] shows low correlations of 0.25 for *Action Consistency* and 0.22 for *Temporal Coherence*. To mitigate this and provide more direct visual evidence, we uniformly sample 40 frames from each video and arrange them into 4×10 grid panels (suppl.). This layout preserves both temporal progression and spatial structure, offering clearer visual evidence. Under this setting, we find a stronger alignment with human judgments: e.g., Gemini-2.5-Pro's correlations with human scores is 0.31 on Action Consistency and 0.26 on Temporal Coherence. Among MLLMs, GPT-5 achieves the highest alignment, with correlations 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. *Action Consistency* (S_{cons}) achieves a correlation of 0.61 while *Temporal Coherence* (S_{temp}) achieves 0.64. Notably, despite being trained on

a much smaller dataset and using only a few features, our metrics outperform GPT-5 by $+\mathbf{35.6}\%$ and $+\mathbf{68.4}\%$ in relative gains, respectively. Both correlations are statistically significant ($p < 0.05$). The 95% bootstrap confidence intervals over the 300 TAG-Bench-v0 videos are $[0.52, 0.63]$ and $[0.59, 0.68]$ for S_{cons} and S_{temp}, respectively, indicating the statistical stability of our metrics.

5.3 Performance on external benchmarks

We evaluate on the Human Anatomy subset of VBench-2.0 [56], which compares videos from four *text-to-video* models (Sora-480p [35], Kling [28], Hunyuan [27], and CogVideo [22]) on a fixed set of text prompts designed to expose anomalies in human appearance and structure in generated videos (e.g, "A woman is cutting objects"). Given these four models, annotators select the more realistic video in pairwise comparisons for each prompt. Following Sec. 5.1, we evaluate only videos with a single visible person per frame, and compute *win ratios* [56]. A model "wins" when its video is preferred by annotators, and its win ratio is the fraction of total comparisons it wins (i.e. number of wins / total pairwise comparisons). Models are ranked by these ratios (higher is better). We then compare these human-derived rankings to those inferred by our S_{temp} metric, by computing win ratio in a similar fashion. Since VBench-2.0 prompts do not correspond to the 10 classes used in training (Sec. 5.1), we evaluate using only S_{temp}, as S_{cons} requires a corresponding action centroid. As shown in Fig. 3, S_{temp} produces the same model ranking as human raters. Notably, the evaluated videos from VBench-2.0 are generated by *text-to-video* models and are prompted with actions not present in our training set (Sec. 5.1). This demonstrates that the learned motion-sensitive embedding generalizes beyond the plausibility learned from actions seen during training.

5.4 Comparing generative models

Having shown that our metrics correlate strongly with human judgments (Sec.5.2), we now use them for a fine-grained comparison of the five state-of-the-art generative models evaluated in our study (Sec.4).

Overall model performance (win ratios). We compare all models on TAG-Bench-v0 using win ratios (Sec. 5.3). As shown in Fig. 3, Wan2.2 [47] achieves the highest win ratios (0.77 for *Action Consistency*, 0.72 for *Temporal Coherence*), outperforming the closed-source Runway Gen-4 [42] (0.65 and 0.61, respectively).

Which actions are difficult to generate for which model? As shown in Fig. 4, Wan2.2 consistently achieves the lowest (best) S_{cons} and S_{temp} for most actions. However, a video may depict the correct action while showing unrealistic motion. For instance, Wan2.2 performs well on "SoccerJuggling" in S_{cons} but relatively poorly in S_{temp}. This gap aligns with human ratings[5], where the human

[5] We report the average metric scores on the original 1–10 scale, computed across valid human raters (Sec. 4) for each video.

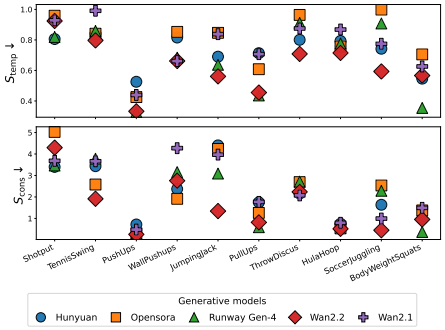


Fig. 4: Comparing generative models. We plot the mean S_{cons} and S_{temp} scores (Sec. 3.3) (lower is better) for each generative model across different actions. Wan2.2 performs best among the other models (low scores in both S_{cons} and S_{temp}). *Shotput* and *JumpingJack* challenge all models, yielding high scores across both metrics.

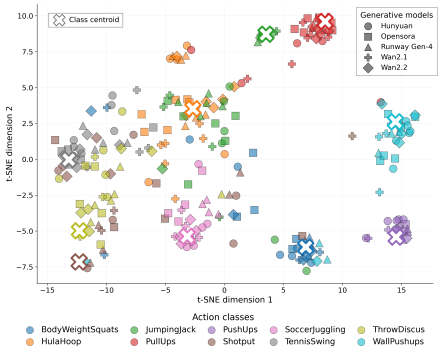


Fig. 5: t-SNE visualization of the embeddings of generated videos along with train centroids. We project the z_{CLS} embeddings of *generated* videos (colored markers) from TAG-Bench-v0 and the corresponding *training* class centroids (white crosses) using t-SNE [33]. Realistic generated videos cluster near their respective class centroids (e.g., Wan2.2 videos for "PullUps", with an average human rating of: **8.41** for *Action Consistency*), while poorly generated videos lie further away (e.g., Wan2.2 videos for "Shotput" with an average human rating of: **4.43**) (Sec. 4).

scores are 8.4 for *Action Consistency* and 7.6 for *Temporal Coherence*, suggesting that viewers perceived the action as semantically correct but noticeably less natural. In contrast, Runway-Gen4 surpasses all models on "BodyWeightSquats" for both metrics. Thus, no single model dominates across all actions.

Are some actions universally harder to generate? From Fig. 4, we note that actions such as "Shotput" and "ThrowDiscus" are challenging across all five models, which show high (i.e., poor) S_{cons} and S_{temp} scores in Fig. 4. This suggests that complex, full-body rotational actions remain a common failure case for current generative models. In contrast, actions involving less dynamic motion such as "PushUps" and "PullUps" are relatively easier to generate, as reflected by low scores in both metrics. This trend is illustrated in the t-SNE [33] visualization in Fig. 5: videos with higher human and automatic ratings (i.e. higher quality) cluster tightly around their class centroids (e.g., "PullUp"), while lower-quality generations appear dispersed away (e.g., "Shotput").

5.5 Analysis

Effect of loss terms. We evaluate the contribution of each loss term in Eq.3 by training variants of the encoder with specific losses removed. From Table 2, we observe that removing the supervised contrastive loss ($\mathcal{L}_{\text{supcon}}$) causes a steep drop in *Action Consistency* (from 0.61 → 0.26), validating that $\mathcal{L}_{\text{supcon}}$ is essential for structuring the embedding space according to action semantics. The addition of $\mathcal{L}_{\text{hard-negative}}$ further refines this embedding space, improving both *Action Consistency* (0.54 → 0.61) and *Temporal Coherence* (0.57 → 0.64). $\mathcal{L}_{\text{supcon}}$ also

$\mathcal{L}_{\text{supcon}}$	$\mathcal{L}_{\text{hard-neg}}$	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.

Visual features	Action Consistency	Temporal Coherence
ViT (TokenHMR)	**0.61**	**0.64**
CLIP	0.51	0.39
DINOv2	0.56	0.38

(a) With human-centric features

Visual features	Action Consistency	Temporal Coherence
CLIP-only features	0.12	0.19
DINOv2-only features	0.14	0.26

(b) Without human-centric features

Table 4: Which visual feature backbone helps the most? (a) Replacing ViT features (from TokenHMR [15]) with CLIP [39] or DINOv2 [37] embeddings of the person cropped from each frame reduces correlation with human judgments. (b) Using CLIP or DINOv2 embeddings alone (without the human-centric 3D or 2D features described Sec. 3.1) as inputs to the encoder performs worst, validating that structured human-centric features are essential for our task.

proves crucial for *Temporal Coherence* (improving scores from 0.38 → 0.64). These results suggest that learning the rules of plausible motion (*how*) is most effective when the embedding space encodes meaningful action semantics (*what*).

Ablation study on input features. To understand the importance of each input feature, we retrain the encoder while zeroing out one input feature at a time (e.g., setting the pose features to zero, while retaining the rest). This retains the full dimensionality of the input and does not alter model capacity. As shown in Table 3, masking any single feature leads to a decrease in performance. For instance, masking $3D$ pose results in a drop in correlation scores for *Action Consistency* from 0.61 → 0.56. Masking all motion features results in the largest drop (*Action Consistency*: 0.61 → 0.46). This highlights the necessity of our multi-feature static and motion representations.

Effect of different visual appearance features. To assess different choices for the visual appearance feature f_{vis}, we train the encoder from scratch by replacing the ViT features from TokenHMR [15] with CLIP [39] and DINOv2 [37] embeddings extracted from the cropped person images (Sec. 5.1), while keeping all other components fixed. As shown in Table 4, replacing ViT features with CLIP or DINOv2 lowers performance (e.g., *Action Consistency*: 0.61 → 0.51, *Temporal Coherence*: 0.64 → 0.39, when replaced with CLIP features). Furthermore, training the encoder using only CLIP or DINOv2 features (i.e., without any human-centric 3D or 2D features) further degrades performance, confirming that our strong results stems not only from the model design and data, but also from the integration of specialized human-centric representations.

Extending TAG-Bench to more action classes. To assess the generalization of our framework, we extend TAG-Bench-v0 from 10 to 23 action classes by incorporating 13 additional actions: *Bowling, Clean&Jerk, GolfSwing, HammerThrow, Hammering, HandStandPushup, JugglingBalls, JumpRope, Lunges, PlayingGuitar, RockClimbingIndoor, RopeClimbing,* and *Surfing*. We follow the

same human evaluation protocol as in Sec. 4 for the additional classes. Our model when trained and evaluated on all 23 classes continues to outperform – correlation scores of *Action Consistency*: 0.59 and *Temporal Coherence*: 0.56, compared to GPT-5's scores of *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 anatomy. We identify a critical gap affecting *all* current benchmarks and metrics for this evaluation. To address this, we introduce **TAG-Bench-v0** and two new metrics, which 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. 5.5 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.

Acknowledgements

This research was supported by Meta. We thank Sai Kumar Dwivedi, Jonathan Granskog, Ed Chien, Robin Kahlow, Lorenzo Torresani, Xingjian (Jessie) Han, and members of our research group at BU for helpful discussions and feedback. We also thank Runway for providing us computational resources for this project.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alison Gopnik, Patricia K. Kuhl, and Andrew N. Meltzoff: The scientist in the crib: Minds, brains, and how children learn (1999)
3. Amazon Web Services, Inc.: Amazon mechanical turk. `https://www.mturk.com/`, accessed: November 2025
4. Animesh, C., Chandraker, M.: Tuned contrastive learning. arXiv preprint arXiv:2305.10675 (2023)
5. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation. arXiv preprint arXiv:2406.03520 (2024)
6. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. arXiv preprint arXiv:2503.06800 (2025)
7. Barker, R.G., Wright, H.F.: Midwest and the USA. Row, Peterson and Company (1954)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? International Conference on Machine Learning (2021)
9. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
10. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
11. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
12. Cho, J., Puspitasari, F.D., Zheng, S., Zheng, J., Lee, L.H., Kim, T.H., Hong, C.S., Zhang, C.: Sora as an agi world model? a complete survey on text-to-video generation. arXiv preprint arXiv:2403.05131 (2024)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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)
14. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
16. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)

18. He, X., Jiang, D., Nie, P., Liu, M., Jiang, Z., Su, M., Ma, W., Lin, J., Ye, C., Lu, Y., et al.: Videoscore2: Think before you score in generative video evaluation. arXiv preprint arXiv:2509.22799 (2025)
19. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., 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)
20. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2021)
21. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems (2022)
22. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arXiv preprint arXiv:2205.15868 (2022)
23. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: Proceedings of the 20th IEEE International Conference on Pattern Recognition (2010)
24. for Intelligent Systems, P.S.D.M.: Smpl made simple faqs, `https://files.is.tue.mpg.de/black/talks/SMPL-made-simple-FAQs.pdf`
25. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
26. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
27. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
28. Kuaishou Technology: Kling: High-fidelity and temporally consistent text-to-video generation. Technical Report (2024), `https://kling.kuaishou.com`
29. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion. arXiv preprint arXiv:2505.01425 (2025)
30. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
31. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. Seminal Graphics Papers: Pushing the Boundaries, Volume 2 (2023)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (2008)
34. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? arXiv preprint arXiv:2501.09038 (2025)
35. OpenAI: Sora: A large-scale diffusion transformer for text-to-video generation. Technical Report (2024), `https://openai.com/research/sora`
36. OpenAI: Gpt-5 system card. Tech. rep., OpenAI (Aug 2025), `https://cdn.openai.com/gpt-5-system-card.pdf`, accessed: 2025-11-10

37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PmLR (2021)
40. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems (2014)
41. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
42. Team, R.R.: Runway gen-4: Advancing realistic text-to-video generation. Technical Report (2024), `https://research.runwayml.com/gen4`
43. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
44. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems (2022)
45. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
46. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
47. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
48. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (2004)
49. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
50. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
51. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models. ACM Computing Surveys (2024)
52. Xue, Z., Luo, M., Grauman, K.: Seeing the arrow of time in large multimodal models. arXiv preprint arXiv:2506.03340 (2025)
53. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)

55. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
56. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
57. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)