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

Xavier Thomas[1] Youngsun Lim[1] Ananya Srinivasan[2*] Audrey Zheng[3*] Deepti Ghadiyaram[1,4†]
[1]Boston University [2]Belmont High School [3]Canyon Crest Academy [4]Runway
{xthomas, youngsun, dghadiya}@bu.edu

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 xthomasbu.github.io/video-gen-evals*

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 [43, 51, 59].

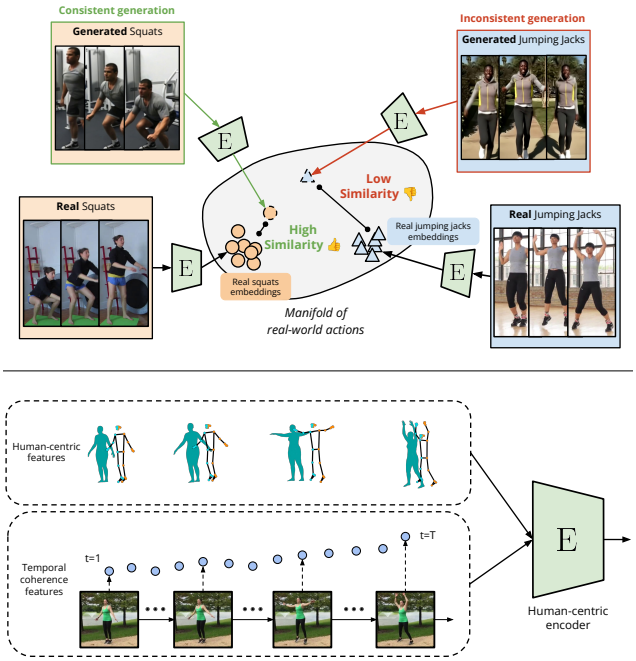

Figure 1. What are the **telltale signs of a generative action video**? We answer this by learning a robust manifold based on appearance and anatomical coherence exhibited by humans performing actions across several real-world videos. This manifold serves as anchors against which we project the features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

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 [21] (e.g., "walking") or procedural (e.g., "making a toast") [8], 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**

*Equal contribution.
†Corresponding author.

coherence of the body movements over time.

Current metrics for judging generated videos, such as pixel-level similarity [28, 61], perceptual quality [55, 68], or text-to-video prompt alignment [25, 38], as well as recent approaches that use MLLMs as judges [6, 23], do not accurately capture the complex motion physics of human actions [15, 52]. 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 [38, 71], 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**). 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** and other benchmarks [71] 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) [69] and text-to-video (T2V) [10] generation settings, and across multiple models. In summary, our contributions are:

- We design a learned latent space that encodes human body geometry, motion smoothness, and temporal coherence to evaluate action quality in generated videos.
- We design **TAG-Bench**, a new 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 [27, 32, 44, 51, 59] has rapidly advanced, producing increasingly photorealistic and temporally coherent content [26, 64]. Beyond visual fidelity, recent work [62] 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 [15, 52]. This highlights the need for metrics that specifically diagnose and quantify these failures.

Distribution and reference based metrics. Metrics such as FVD [55] 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 [28], SSIM [61], and LPIPS [68] assess visual similarity to reference frames but ignore temporal coherence critical to video perception. Recent efforts like Physics-IQ [43] 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 [25] 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 [23] predicts fine-grained human ratings across dimensions such as visual quality and temporal consistency, while VideoPhy [6] 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 [38] provides a large-scale framework for assessing visual quality, motion quality, temporal consistency, and text–video alignment. VBench2.0 [71] extends this with finer-grained anatomy-related criteria. However, none explicitly assess whether human actions are executed plausibly over time. Our proposed TAG-Bench addresses this gap through targeted met-

rics 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 multiple human-centric features encompassing appearance, skeletal geometry, and motion dynamics detailed below.

3.1.1. 3D features

We employ Skinned Multi-Person Linear (SMPL) [39], 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 [39], capturing how the human body moves during a particular action. β captures shape-specific characteristics [39], such as overall build (e.g., tall vs. short) and limb proportions. go describes the global body rotation computed from the pelvis joint [20], 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) [30] models.

Motivation to use SMPL: We believe that the detailed 3D features are essential for capturing the complex kinematics of a human body in action. SMPL features are invariant to appearance and scene context [31, 39], and focus specifically on the geometry of the human body. Thus, they serve as powerful ingredients to assess if generated humans follow the same physical dynamics as real-world humans.

3.1.2. 2D features

While 3D representations capture detailed joint rotations and body shapes, SMPL is trained solely on real human data [39], constraining its parameters to remain within anatomically plausible body configurations [47]. While this is an excellent prior to learn human action dynamics in *real* videos, it may overlook anomalies such as elongated limbs or implausible joint configurations common in *generated* videos. Thus, we also incorporate 2D joint keypoints [11] (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.

3.1.3. Visual appearance features

Although 3D and 2D features capture body structure, they are inherently and intentionally appearance-invariant [11, 39, 48]. To complement them, we extract visual appearance features (f_{vis}) using pre-trained image-based visual backbones (e.g., ViT [18]). These features capture cues such as clothing, color, and action-relevant objects, all of which influence how an action is visually perceived.

3.1.4. 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 artifacts common in generative media 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 in real videos.

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

3.2.1. Encoder architecture and training

Constructing temporal windows: 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 $t \in$

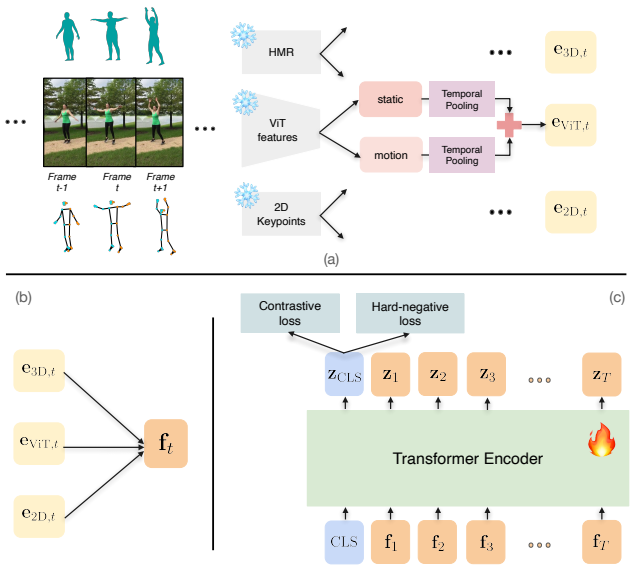

Figure 2. **Architectural overview of the encoder we train to learn the real-world action manifold.** We extract per-frame static human-centric and temporal motion features (Fig. (a)) (Sec. 3.1), and aggregate them, yielding one embedding for each frame (Fig. (b)) (Sec. 3.2.1). We prepend a [CLS] token 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.

$\{1, \ldots, T\}$ 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 Simonyan and Zisserman [49] 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 enable the model to capture the local temporal evolution of each input feature (e.g., body pose θ).

We then combine the static and motion encodings for each frame t via element-wise addition, similar to the additive fusion used in GENMO [35]. This way, the motion pathway provides residual information to the otherwise static state:

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

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 [57] between a learnable query vector $\mathbf{q} \in \mathbb{R}^d$ and a linear projection of each input feature using $\mathbf{W}_a \in \mathbb{R}^{d \times d}$. The softmax operation ensures that the weights are positive and sum to 1. This attention mechanism is learned jointly with the rest of 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] [9], 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.

3.2.2. Training objective

Our goal is to learn an effective latent space of real-world human actions. We achieve this via a multi-loss objective that combines two complementary goals:

(i) Learn action semantics: We use a supervised contrastive loss ($\mathcal{L}_{\text{supcon}}$) [5] 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 contribution of the two loss terms. This combined objective encourages the encoder to capture both action semantics (*what* action is performed) 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 ($S_{\text{cons}} = \left\|\mathbf{z}^{\text{gen}}_{\text{video}} - \mathbf{c}_k\right\|_2$) 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, thereby complementing each other. 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 [24, 71] 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". TAG-Bench comprises videos generated from several open- and closed-source image to video 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 [50] 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*. For each action, we sample 6 videos at random and extract their first frame. This frame serves as the input to image-to-video (I2V) generation models – Wan2.1 [59], Wan2.2 [59], Hunyuan [32], Opensora [72], and Runway Gen-4 [51] – 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 [4]. 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 (Sec 5.2) and external benchmarks (Sec 5.3). We then leverage our metrics to 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 [19] (Sec. 3.1.1) and 2D keypoints using DW-Pose [67] (Sec. 3.1.2). TokenHMR employs Detectron2 [63] to obtain bounding boxes for the person; the cropped image is then used to infer SMPL parameters (Sec. 3.1.1). Visual appearance features (Sec. 3.1.3) are ob-

tained from the frozen ViT-H/16 backbone of TokenHMR. Motion features (Sec. 3.1.4) 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. 3.2.1), 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 [22]. 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 (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 [40] 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 [23, 24], 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 (Sec. 4), including feature-based methods, fine-tuned MLLM 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 PIQUE [58], CLIP-sim [48], and CLIP-Score [25] fail to capture motion dynamics, correlating poorly with human ratings (< 0.22). TRAJAN [3], which tracks global scene smoothness via point trajectories, also underperforms, indicating that overall scene coherence does not imply accurate action. The metrics from VBench-2.0 [71] designed to assess human anatomy and identity per frame also show very low correlations on our benchmark (< 0.11).

Method	Corr. with Action Consistency ↑	Corr. with Temporal Coherence ↑
Random	-0.07	-0.11
Feature-based automatic metrics		
PIQUE [58]	-0.19	-0.13
BRISQUE [42]	-0.04	0.01
CLIP-sim [48]	0.03	0.16
DINO-sim [13]	0.08	0.21
SSIM-sim [61]	-0.08	-0.04
MSE-dyn [60]	-0.15	-0.08
SSIM-dyn [60]	-0.06	-0.03
CLIP-Score [25]	0.08	0.00
X-CLIP-Score [41]	0.00	-0.07
TRAJAN [3]	-0.12	-0.12
VideoMAE(UCF101)-classification	0.18	0.17
VBench-2.0 [71] (Human Anatomy)	-0.40	0.02
VBench-2.0 [71] (Human Identity)	0.06	0.02
VBench-2.0 [71] (Human Clothes)	0.12	0.11
MLLM-based fine-tuned metrics		
🔓 VideoScore [23] (Visual Quality)	-0.12	-0.06
🔓 VideoScore [23] (Temporal Consistency)	-0.09	-0.04
🔓 VideoScore [23] (Dynamic Degree)	-0.19	-0.16
🔓 VideoScore [23] (T2V Alignment)	-0.07	-0.04
🔓 VideoScore [23] (Factual Consistency)	-0.14	-0.08
🔓 VideoScore2 [24] (Visual Quality)	0.14	0.16
🔓 VideoScore2 [24] (T2V Alignment)	0.17	0.09
🔓 VideoScore2 [24] (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 [36]	-0.17	-0.14
🔓 LLaVA-v1.6-mistral-7b-hf [37]	-0.10	0.18
🔓 Idefics2-8B [34]	-0.05	-0.06
🔓 Qwen3-VL-8B-Instruct [66]	0.34	0.28
🔒 Gemini-2.5-Flash [16]	0.40	0.25
🔒 Gemini-2.5-Pro [16]	0.31	0.26
🔒 GPT-4o [1]	0.34	0.31
🔒 GPT-5 [45]	<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 [54] 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.

MLLM-based fine-tuned evaluators: Recent methods like VideoScore [23, 24] and VideoPhy-2 [7] 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

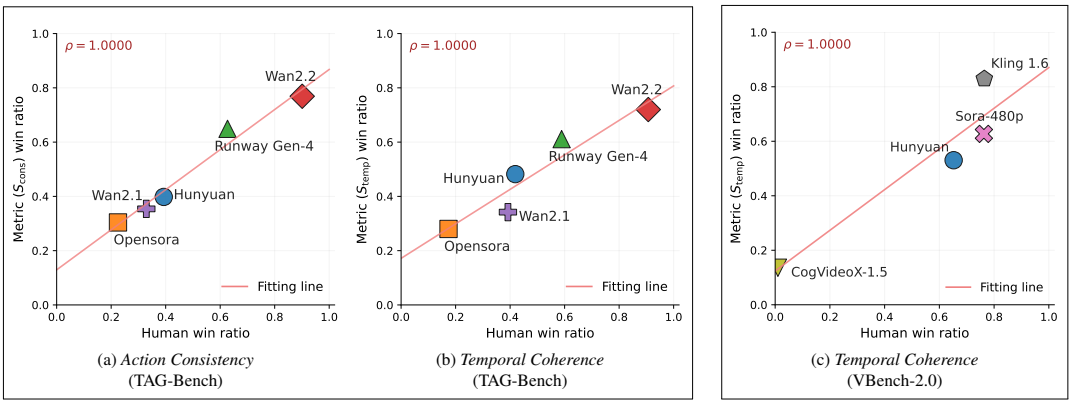

Figure 3. **Model comparisons on TAG-Bench and VBench-2.0 Human Anatomy.** We compare models pairwise for the same input prompt; for each pair, the model with the higher score (human or metric) is the winner. We then plot the win ratios (see Sec. 5.3) of human scores (x-axis) against win ratios from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

align with human judgments. Directly using raw video inputs proved unreliable, as MLLMs often fail to capture fine-grained temporal details in human motion [65]. For instance, Gemini-2.5-Pro [16] 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 (example in Appendix). This layout preserves both temporal progression and spatial structure, offering clearer visual evidence. Under this setting, we find a stronger alignment with human judgments: for example, 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: Compared to all methods in Table 1, 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.

5.3. Performance on external benchmarks

We evaluate on the Human Anatomy subset of VBench-2.0 [71], which compares videos from four *text-to-video* models (Sora-480p [44], Kling [33], Hunyuan [32], and CogVideo [27]) 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* [71]. 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 (Sec. 4) using win ratios (Sec. 5.3). As shown in Fig. 3, the open-source Wan2.2 [59] achieves the highest win ratios (0.77 for *Action Consistency*, 0.72 for *Temporal Coherence*), outperforming the closed-source Runway Gen-4 [51] (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 poorly in S_{temp}. This gap aligns with human ratings[1], where the

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

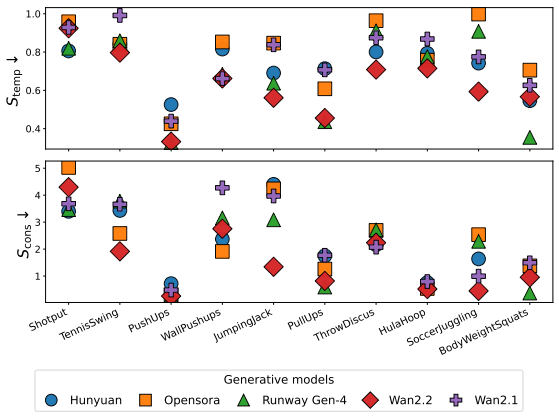

Figure 4. **Comparing generative models.** We plot the mean S_{cons} and S_{temp} scores (Sec. 3.3) (lower is better) for each generative model across different actions. Wan2.2 performs best among the other models (low scores in both S_{cons} and S_{temp}). *Shotput* and *JumpingJack* challenge all models,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

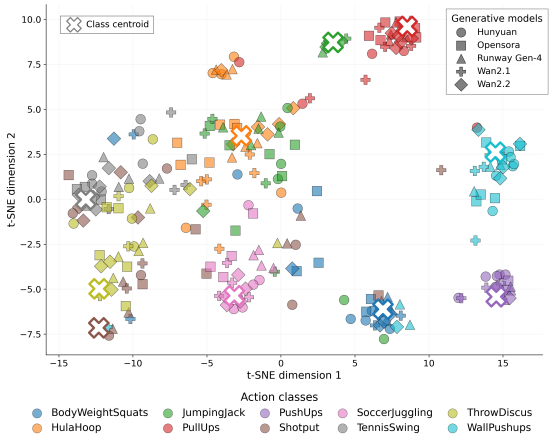

Figure 5. **t-SNE visualization of the embeddings of generated videos along with train centroids.** We project the z_{CLS} embeddings of *generated* videos (colored markers) from TAG-Bench and the corresponding *training* class centroids (white crosses) using t-SNE [56]. Realistic generated videos cluster near their respective class centroids (e.g., Wan2.2 videos for "PullUps", with an average human rating of: $\mathbf{8.41}$ for *Action Consistency*), while poorly generated videos lie further away (e.g., Wan2.2 videos for "Shotput" with an average human rating of: $\mathbf{4.43}$) (See Sec. 4).

human 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 Wan2.2 and other models on "BodyWeightSquats" for both metrics. Thus, no single model dominates across all actions.

Are some actions universally harder to generate? We examine whether certain actions are consistently more difficult for generative models. Notably, actions such as "Shotput" and "ThrowDiscus" are challenging across all five

$\mathcal{L}_{\text{supcon}}$	$\mathcal{L}_{\text{hard-neg}}$	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. 3.1.4). Removing motion causes the largest degradation.

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 repetitive and 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 [56] visualization in Fig. 5: videos with higher human and automatic ratings (i.e. higher quality) cluster tightly around their class centroids (e.g., "PullUp" videos across models), while lower-quality generations appear further away (e.g., "Shotput" videos across all models).

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 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*). Together, both losses are essential for capturing action semantics and temporal coherence.

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 re-

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 [19]) with CLIP [48] or DINOv2 [46] 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 evaluating realism of human actions.

tains 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 $0.61 \rightarrow 0.56$. Masking all motion features results in the largest drop (*Action Consistency*: $0.61 \rightarrow 0.46$). 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 [19] with CLIP [48] and DINOv2 [46] 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*: $0.61 \rightarrow 0.51$, *Temporal Coherence*: $0.64 \rightarrow 0.39$, 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.

Training on more data: We test whether additional training data improves the learned embedding space by augmenting UCF101 [50] with Kinetics-700 [14] clips depicting the same action classes. As Kinetics videos are not temporally trimmed to only showcase the action [14], we adopt an active sampling approach: our UCF101-trained model (Sec. 5.1) scores each temporal window from Kinetics-700 by its distance to the corresponding class centroid. By retaining only the most representative 20% windows (i.e., with least distances) and augmenting the training set with these samples, correlation scores improve from $0.61 \rightarrow 0.65$ for *Action Consistency* and $0.64 \rightarrow 0.65$ for *Temporal Coherence*. However, performance drops when incorporating samples beyond the top 20% (i.e., less representative windows farther from the class centroid). For example, using the top 30% reduces *Action Consistency* from $0.65 \rightarrow 0.63$, and reduces *Temporal Coherence* from $0.65 \rightarrow 0.63$.

6. Discussion and Future Work

In this work, we evaluate generated human actions 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** and two new metrics, which demonstrate strong alignment with human perceptual judgment and generalizability across diverse generation models. Future work will extend this framework to longer-form videos and explore integrating our 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. 6

[2] Patricia K. Kuhl Alison Gopnik 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. 6, 8

[4] Amazon Web Services, Inc. Amazon mechanical turk. https://www.mturk.com/. Accessed: November 2025. 5

[5] Chaitanya Animesh and Manmohan Chandraker. Tuned contrastive learning. *arXiv preprint arXiv:2305.10675*, 2023. 4

[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. 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. 6, 8

[8] Roger G Barker and Herbert F Wright. *Midwest and the USA*. Row, Peterson and Company, 1954. 1

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *International Conference on Machine Learning*, 2021. 4

[10] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[11] 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. 3

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision*, 2021. 8

[13] 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. 6

[14] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 9

[15] 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. 2, 5

[16] 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. 6, 7

[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 8

[18] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[19] 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. 5, 9, 4, 7

[20] Perceiving Systems Department MPI for Intelligent Systems. Smpl made simple faqs. https://files.is.tue.mpg.de/black/talks/SMPL-made-simple-FAQs.pdf. 3

[21] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, 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. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 6

[23] 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. 2, 6, 8

[24] 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. 5, 6, 8

[25] 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. 2, 6, 8

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 2022. 2

[27] 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. 2, 7, 5

[28] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *Proceedings of the 20th IEEE International Conference on Pattern Recognition*, 2010. 2

[29] International Telecommunication Union. Methodologies for the subjective assessment of the quality of television pictures. Technical report, ITU Radiocommunication Sector (ITU-R), 2019. 2

[30] 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. 3

[31] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[32] 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. 2, 5, 7, 9

[33] Kuaishou Technology. Kling: High-fidelity and temporally consistent text-to-video generation. *Technical Report*, 2024. https://kling.kuaishou.com. 7, 5

[34] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 2024. 6

[35] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025. 4

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 2023. 6

[37] 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. 6

[38] 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. 2

[39] 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. 3

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[41] 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. 6, 8

[42] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012. 6, 8

[43] 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, 2

[44] OpenAI. Sora: A large-scale diffusion transformer for text-to-video generation. *Technical Report*, 2024. https://openai.com/research/sora. 2, 7, 5

[45] OpenAI. Gpt-5 system card. Technical report, OpenAI, 2025. Accessed: 2025-11-10. 6

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 9

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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. 3

[48] 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, 6, 9, 8

[49] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 2014. 4

[50] 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. 5, 9

[51] Runway Research Team. Runway gen-4: Advancing realistic text-to-video generation. *Technical Report*, 2024. https://research.runwayml.com/gen4. 1, 2, 5, 7

[52] 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. 2

[53] Xavier Thomas and Deepti Ghadiyaram. What's in a latent? leveraging diffusion latent space for domain generalization. *arXiv preprint arXiv:2503.06698*, 2025. 6

[54] 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. 6

[55] 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. 2

[56] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 8, 5, 6

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[58] 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. 6, 8

[59] 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, 2, 5, 7

[60] 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. 6, 8

[61] 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. 2, 6, 8

[62] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025. 2

[63] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5, 4

[64] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2024. 2

[65] Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025. 7

[66] 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. 6

[67] 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. 5

[68] 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. 2

[69] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[70] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 8

[71] 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. 2, 5, 6, 7, 8

[72] 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. 5

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

Supplementary Material

Table of Contents

A. Human Evaluation User Interface (UI)
B. TAG-Bench: Model configurations, subject rejection, aggregation of opinion scores
- B.1 Model configurations
- B.2 Subject rejection
- B.3 Convergence analysis of human ratings

C. Additional implementation details
- C.1 TokenHMR feature extraction
- C.2 Training data processing
- C.3 Human-centric features extracted from each frame

D. Win Ratios (TAG-Bench, VBench-2.0)
E. Additional Experiments
- E.1 Attention weights

F. Baselines
- F.1 Feature-based metrics
- F.2 Multi-modal Large Language Models (MLLMs) based metrics

G. MLLM Prompting
- G.1 Prompt used for MLLM evaluation
- G.2 Limited impact of in-context learning

A. Human Evaluation User Interface (UI)

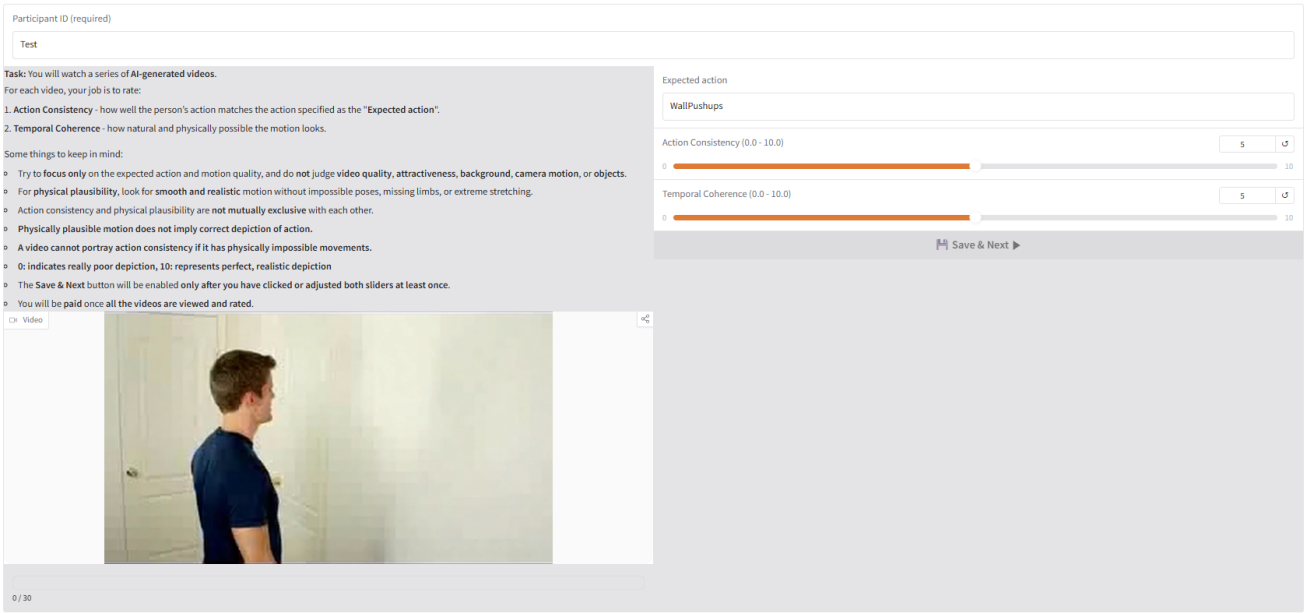

Figure 6. **User interface used for the human evaluation study.** Participants were asked to rate each AI-generated video along two dimensions: *Action Consistency* (how accurately the motion matches the described action) and *Temporal Coherence* (how natural and physically realistic the motion appears). Each participant viewed 30 videos and provided ratings on a scale from 0 to 10.

B. TAG-Bench: Model configurations, Subject rejection, Aggregation of Opinion Scores

B.1. Model configurations

Below are the configurations of the image-to-video (I2V) models used to generate the videos evaluated in TAG-Bench.

Model	Resolution	Model name
Wan2.1 [59]	1104×816	Wan2.1-I2V-14B-720P[2]
Wan2.2 [59]	1280×720	Wan2.2-I2V-A14B[3]
Opensora [44]	1024×576	Opensora-768px[4]
Hunyuan [32]	1088×832	HunyuanVideo-I2V-720p[5]
Runway Gen-4 [51]	1280×720	Gen4-Turbo[6]

Table 5. **Image-to-video (I2V) models used in TAG-Bench.** Resolution indicates the native output resolution for each model variant.

B.2. Subject rejection

We also select only workers with a Human Intelligence Task (HIT) approval rate of $>= 99\%$. Furthermore, to ensure that the subjective ratings are consistent and reliable, we apply a three-stage filtering process before computing the final Mean Opinion Scores (MOS) and comparing them with model predictions, described next.

(1) Repeated-video consistency filtering Among the 30 videos shown to an annotator, 5 are intentionally duplicated to verify response consistency; if the ratings on the duplicated videos are inconsistent, we exclude those participants. Specifically, for each participant, we compute the standard deviation of their ratings across the repeated samples. We retain only those participants whose scores fall within the 95th percentile on the repeated videos. This step helps remove raters giving inconsistent scores for identical stimuli, resulting in retaining 207 raters of the total 246 participants.

(2) Subject rejection. Next, we apply the subject rejection procedure described in [29] to further eliminate statistically unreliable participants. The method evaluates each participant's deviation from the population mean using two statistics:

$$R_1 = \frac{P_i + Q_i}{N_i}, \quad R_2 = \frac{|P_i - Q_i|}{P_i + Q_i}, \tag{5}$$

where P_i and Q_i are the counts of a participant's scores that lie respectively above or below the population mean by more than the threshold (either $2S$ or $\sqrt{20}S$, depending on the distribution's kurtosis), and N_i is the number of videos rated. Participants with $R_1 > 0.05$ and $R_2 < 0.3$ were rejected, as were those who rated fewer than ten videos. This step resulted in rejecting 14 additional raters, leaving **193 valid participants**.

(3) Inter-rater reliability filtering. Finally, we assess inter-rater reliability by computing the Spearman correlation coefficient (ρ) between each participant's ratings and the aggregated mean ratings of the remaining participants. To remove inconsistent evaluations, we excluded raters with $\rho < 0.55$; the cutoff was set empirically to balance reliability and participant retention, resulting in 121 retained raters for *Action Consistency* and 141 for *Temporal Coherence*.

After completing the three filtering stages, we computed the **MOS** for each video along both evaluation axes—*Action Consistency* and *Temporal Coherence*—as well as their **z-score normalized** versions:

$$z_i = \frac{x_i - \mu}{\sigma}, \tag{6}$$

where x_i is the raw MOS of video i, μ is the mean MOS across all videos, and σ is the standard deviation. We use these as final scores throughout our work and consistency analyses with the model evaluation results.

[2] https://github.com/Wan-Video/Wan2.1
[3] https://github.com/Wan-Video/Wan2.2
[4] https://github.com/hpcaitech/Open-Sora
[5] https://github.com/Tencent-Hunyuan/HunyuanVideo-I2V
[6] https://runwayml.com/research/introducing-runway-gen-4

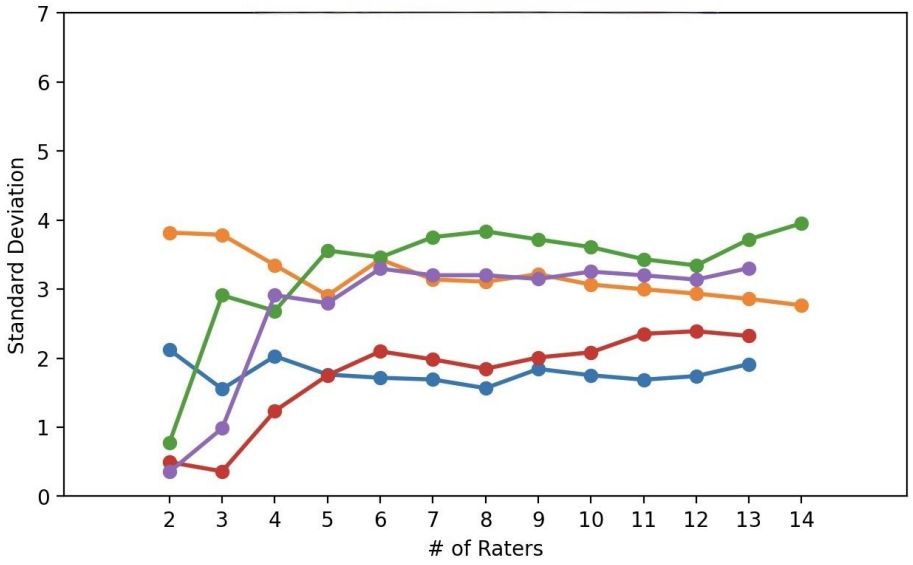

Figure 7. Standard deviation vs. number of raters for five randomly selected videos (*Action Consistency*). Each curve shows the evolution of the MOS standard deviation as more raters are included. Notice how the standard deviation is stabilizing as we add more raters.

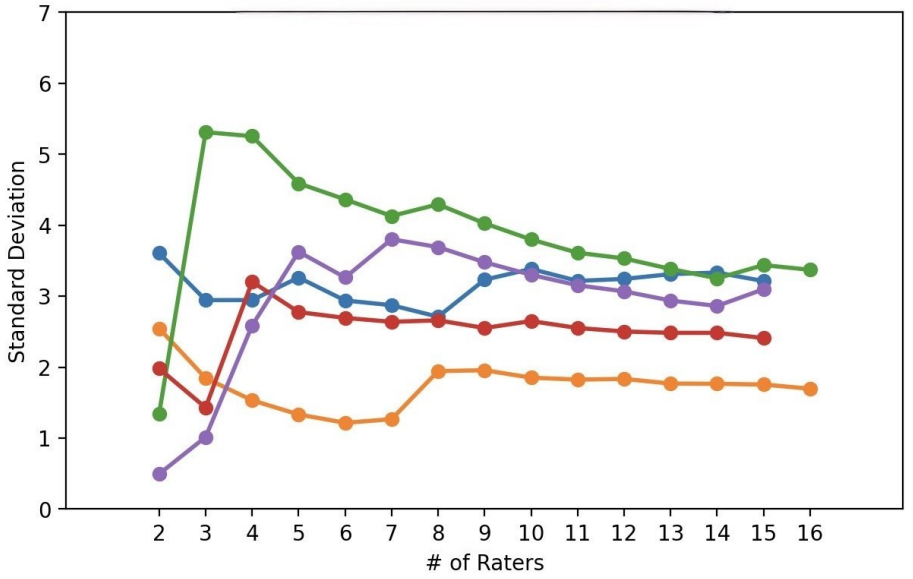

Figure 8. Standard deviation vs. number of raters for five randomly selected videos (*Temporal Coherence*). Each curve shows the evolution of the MOS standard deviation as more raters are included. Notice how the standard deviation is stabilizing as we add more raters.

B.3. Convergence analysis of human ratings

To ensure the reliability of our human evaluations, we examined the stability of the standard deviation of scores as the number of raters increased. For each evaluation axis (*Action Consistency* and *Temporal Coherence*), we selected the five videos evaluated by the largest number of participants and computed the standard deviation of their Mean Opinion Scores (MOS) in a cumulative manner, progressively adding more raters.

As shown in Figures 7 and 8, the standard deviation stabilizes as the number of raters increases, typically converging after about 9–10 participants. This convergence indicates that the variability in human judgments remains bounded, suggesting that our collected human evaluations are statistically stable and reliable. Hence, the aggregated MOS scores used in our main experiments are based on a sufficiently large and consistent set of raters.

C. Additional implementation details

C.1. TokenHMR feature extraction

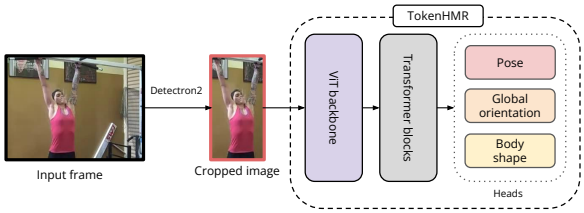

Figure 9. **TokenHMR-based feature extraction.** Each input frame is processed by Detectron2 [63] to obtain a bounding box of the person, which is cropped and passed to TokenHMR [19]. TokenHMR uses a ViT-H/16 backbone, followed by transformer blocks and task-specific heads, to predict SMPL parameters: pose (θ), global orientation (go), and body shape (β) in Sec. 3.1.1. Intermediate features from the ViT backbone are used as visual appearance features (f_{vis}) in Sec. 3.1.3. We adapt the figure from TokenHMR [19].

C.2. Training data processing

We use Detectron2 [63] to count the number of people in each frame and discard videos containing more than one person in any frame, as mentioned in Sec. 5.1. To ensure input features correspond only to the person performing action, we retain videos with only a single visible person. This filtering step yields 930 videos out of the $1,279$ videos across the 10 selected classes (Sec. 4).

C.3. Human-centric features extracted from each frame

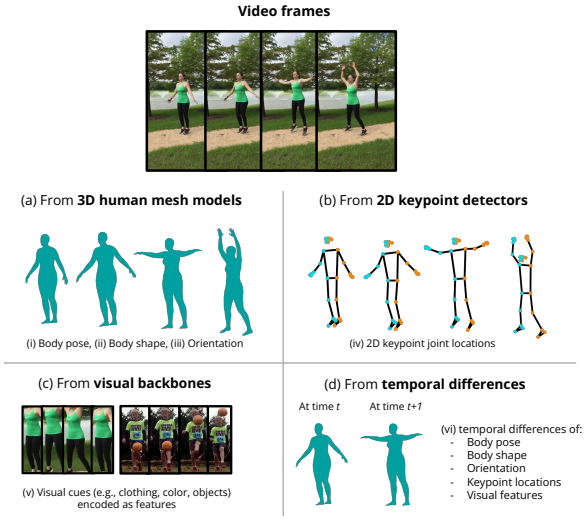

Figure 10. **Human-centric input features.** From each frame, we extract (a) 3D pose, body shape, and global orientation (Sec. 3.1.1), (b) 2D keypoints (Sec. 3.1.2), and (c) visual appearance features (Sec. 3.1.3) to describe the body's state in that frame. (d) We additionally compute temporal differences of each feature to capture frame-to-frame motion dynamics (Sec. 3.1.4)

Feature	Pose	Global orientation	Body shape	Keypoints	Visual features
Dimension	23×3×3	1×3×3	10	60×2	1024

Table 6. **Per-frame input feature dimensions.** Pose and Global orientation are 3×3 rotation matrices representing joint rotations (23 joints) and the global orientation (pelvis joint), respectively. Body shape is a 10D vector. Keypoints denote 2D keypoints comprising 18×2 body and 42×2 hand coordinates. Visual features are features extracted from the ViT backbone of TokenHMR [19]. All features are flattened and normalized across the dataset before being input to the encoder.

D. Win Ratios (TAG-Bench, VBench-2.0)

We show the win ratios obtained in Sec. 5.3 (VBench-2.0 [71]) and Sec. 5.4 (TAG-Bench) below.

Model	Human win ratio	Metric win ratio
Sora-480p [15]	0.76	0.63
Kling 1.6 [33]	0.78	0.83
Hunyuan [32]	0.65	0.53
CogVideoX-1.5 [27]	0.00	0.14

Table 7. **Win ratios for *Temporal Coherence* on the VBench-2.0 human-anatomy subset**: As detailed in Sec. 5.3, the ranking of models inferred by our *Temporal Coherence* (S_{temp}) metric aligns with human judgment.

Model	Human win ratio	Metric win ratio
Hunyuan [32]	0.39	0.40
Opensora [44]	0.23	0.30
Runway Gen-4 [51]	0.63	0.65
Wan2.2 [59]	0.90	0.77
Wan2.1 [59]	0.33	0.35

Table 8. **Win ratios for *Action Consistency* on TAG-Bench**: As detailed in Sec. 5.4, the ranking of models inferred by our *Action Consistency* (S_{cons}) metric aligns with human judgment.

Model	Human win ratio	Metric win ratio
Hunyuan [32]	0.42	0.48
Opensora [44]	0.17	0.28
Runway Gen-4 [51]	0.59	0.61
Wan2.2 [59]	0.91	0.72
Wan2.1 [59]	0.39	0.34

Table 9. **Win ratios for *Temporal Coherence* on TAG-Bench**: As detailed in Sec. 5.4, the ranking of models inferred by our *Temporal Coherence* (S_{temp}) metric aligns with human judgment.

E. Additional Experiments

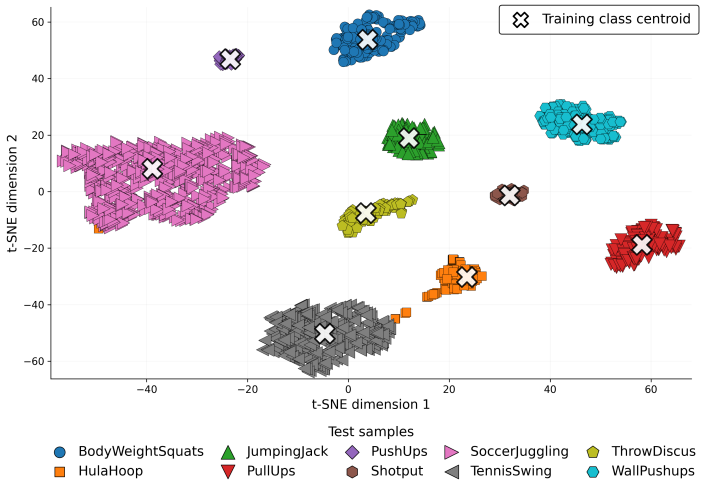

Figure 11. **t-SNE visualization of the embeddings of unseen test videos of diverse actions.** We project the z_{CLS} embeddings of *unseen real* test videos (colored markers) and the corresponding *training* class centroids (crosses) using t-SNE [56]. It is evident that unseen test videos cluster around their respective class centroids, indicating that the learned embedding space captures compact and semantically meaningful action structure.

Action class separability. We evaluate whether the learned embedding space places unseen real videos into the correct action-specific regions. For this, we extract the temporal window-level embeddings (z_{CLS}) (Sec. 3.2) from the held-out test set, run K-means clustering on these embeddings with $K{=}10$ (matching the number of action classes (Sec. 4)), and measure alignment of the cluster assignments with the ground-truth labels using Normalized Mutual Information

(NMI) [53]. We obtain a high NMI score of 0.98, indicating that unseen videos map to the correct action-specific region in the embedding space. Figure 11 visualizes this separation using t-SNE [56]: we plot embeddings of test windows along with the corresponding class centroids computed from the training set. Each action forms a tight, well-separated cluster around its real-video centroid. This shows that the embedding space captures a generalizable notion of action semantics, and that class centroids derived from real videos serve as reliable reference points for evaluating generated videos.

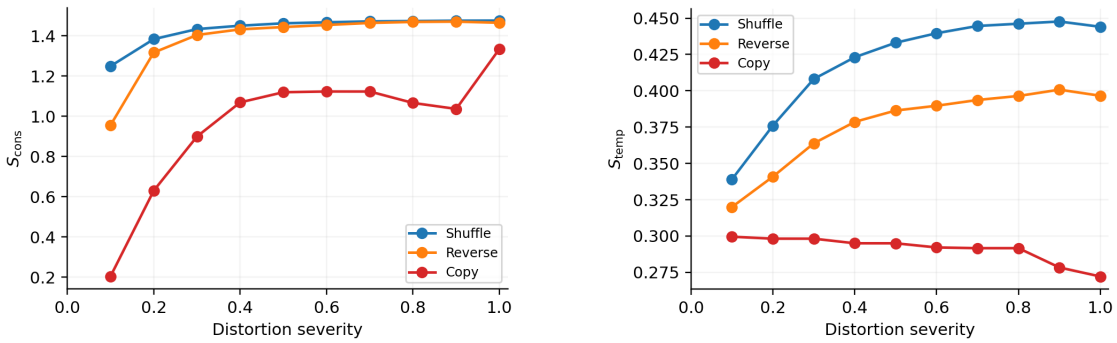

Figure 12. Sensitivity of *Action Consistency* (S_{cons}) and *Temporal Coherence* (S_{temp}) to controlled temporal distortions. The mean value across all test samples is plotted for each distortion type and severity.

Sensitivity to temporal distortions. We measure whether the learned embedding space and metrics are temporally sensitive. For this, we apply controlled temporal corruptions to unseen real videos as done on training data (Sec. 3.2.2). We describe each perturbation below:

- **Shuffle**: randomly reorders a portion of frames, breaking local motion continuity. Severity controls the number of shuffled frames: 0.1 shuffles 10% of frames, while 1.0 shuffles all frames in the window.
- **Reverse**: reverses the frame order in a temporal segment of the window, inverting the temporal direction of motion. Severity determines the length of the reversed segment: at 0.1, only 10% of frames are reversed, while 1.0 reverses the entire window.
- **Copy**: replaces a contiguous segment of frames with copies of the first frame. Severity specifies how much of the sequence is replaced: 0.1 replaces 10% of the frames, whereas 1.0 replaces all frames with the first frame of the window.

We then compute both S_{cons} and S_{temp} on these perturbed videos. As shown in Fig. 12, both metrics monotonically increases (where lower is better, as described in Sec. 3.3) with increasing distortion severity: S_{cons} increases as corrupted videos drift away from the real-action manifold, and S_{temp} increases as temporal smoothness deteriorates (with the exception of "Copy", due to identical adjacent frames resulting in low differences in their frame embeddings (Sec. 3.3)). This confirms that the learned space is explicitly sensitive to temporal dynamics and does not rely solely on visual appearance at the frame level.

Effect of temporal window length We train separate models using different temporal window sizes ($T \in \{4, 8, 16, 32, 64, 128, 256\}$). For videos shorter than T frames, the final frame is repeated to match the window length. As shown in Table 10, performance improves significantly when increasing T from 4 to 32 (0.39 $\rightarrow$ 0.61), for *Action Consistency*), indicating the importance of sufficient temporal context. However, increasing T beyond 32 yields no further gains while substantially increasing computational cost. Thus, we adopt T=32 for all experiments.

Window size (T)	4	8	16	32	64	128	256
Action Consistency	0.39	0.46	0.59	**0.61**	0.56	0.57	0.55
Temporal Coherence	0.43	0.43	0.60	**0.64**	0.58	0.61	0.59

Table 10. **Effect of temporal window size.** Increasing the temporal window length T improves performance significantly up to T=32, beyond which no gains are observed, while computational cost rises sharply.

E.1. Attention weights

We visualize the attention weights from our embedding model (Sec. 3.2) in Fig. 13 and Fig. 14. Figure 13 shows the average attention distribution over all real test samples, while Figure 14 breaks this down by action class. We observe that attention weights vary by action class. Although the visual features (ViT features from TokenHMR [19]) dominate overall, these features inherently encode both appearance cues and implicit geometric structure (see Fig. 9), as they are used to infer the SMPL parameters (Sec. 3.1.1). The model also assigns high weight to 3D pose features, indicating that anatomically grounded signals are essential for modeling human action. For certain actions, the model adapts its focus toward other input features, for example, the "HulaHoop" class shows increased reliance on global orientation.

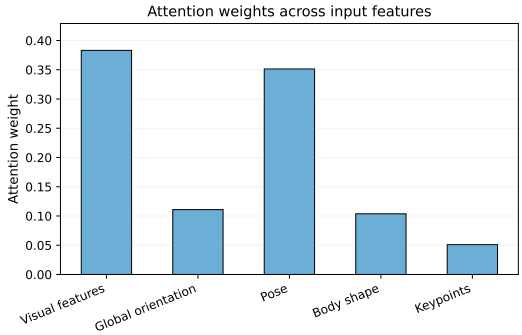

Figure 13. Average attention weights averaged over all real test videos.

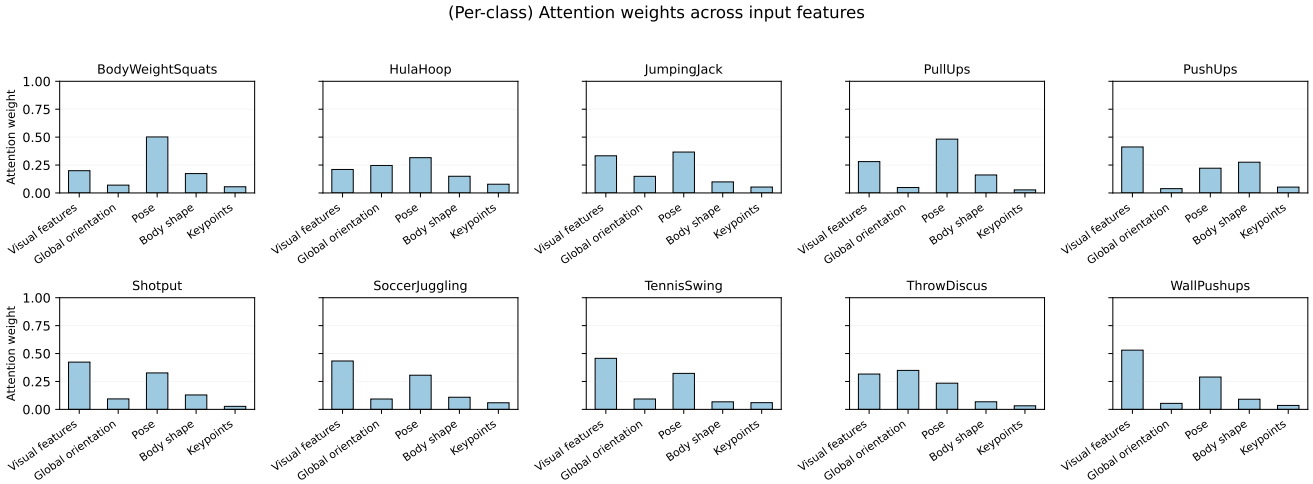

Figure 14. Per-class attention weights, averaged over all real test videos.

F. Baselines

We describe the baselines we report in Table 1 below.

F.1. Feature-based Metrics

TRAJAN [3]. Evaluates motion realism by reconstructing point trajectories using a video autoencoder; inconsistent or implausible motion yields lower Average Jaccard scores (0–1), with higher scores indicating more natural and realistic videos.

PIQUE [58]. A no-reference image quality metric that estimates distortion by analyzing local blocks in each frame and scoring only perceptually significant regions; outputs a 0–100 distortion score, where lower values indicate better visual quality.

BRISQUE [42]. A no-reference quality metric that detects deviations from natural image statistics on a per-frame basis, producing a 0–100 distortion score (lower is better).

SSIM similarity [61]. Computes frame-wise structural similarity between generated and reference videos and averages scores across frames (0–1, higher is better).

CLIP Similarity. Computes the average cosine similarity of CLIP VIT-B/32 [48] embeddings for adjacent frames.

DINO Similarity. Computes the average cosine similarity of DINO ViT-B/16 [12] embeddings for adjacent frames.

MSE Dyn. [60]. Computes the average mean squared error of every fourth frame.

SSIM [60] Dyn. Computes the structural similarity of every fourth frame based on luminance, contrast, and spatial arrangement.

CLIP Score [25] Computes the average cosine similarity of the CLIP ViT-B/32 [48] embedding of a given text prompt and the CLIP embeddings of each frame. Each text prompt follows the format "A person doing [action]."

X-CLIP Score. Computes the cosine similarity of X-CLIP [41] embeddings of a given text prompt and X-CLIP video embeddings. The text prompts each follow the format "A person doing [action]."

VBench-2.0 [71] (Human Anatomy). Evaluates frame-level anatomical correctness by detecting abnormalities in body structure, hands, and faces using three anomaly-detection models; the final score is the percentage of frames without detected anomalies.

VBench-2.0 [71] (Human Face). Measures whether the same person is preserved across frames by computing facial feature similarity (ArcFace [17]) relative to the first frame, ignoring segments with multiple or no people.

F.2. Multi-modal Large Language Models (MLLMs) based metrics

VBench-2.0 [71] (Human Clothes). Assesses whether the person's clothing remains consistent across the video by probing an MLLM (LLaVA-video-7B [70]).

Videophy2 [7] Measures the physical commonsense and semantic adherence of a video. Physical commonsense captures if physical rules (e.g. gravity, collision dynamics) are obeyed. Semantic adherence measures the adherence of a video to its text description.

Videoscore [23]. Evaluates generated videos across five dimensions: visual quality, temporal consistency, dynamic degree, text–video alignment, and factual consistency, using a model trained on large-scale human ratings; each dimension is scored from 1 to 4.

Videoscore2 [24]. Evaluates videos along visual quality, text–video alignment, and physical consistency using a model trained on large-scale human ratings; outputs scores from 1 to 5 along with reasoning for each dimension.

G. MLLM Prompting

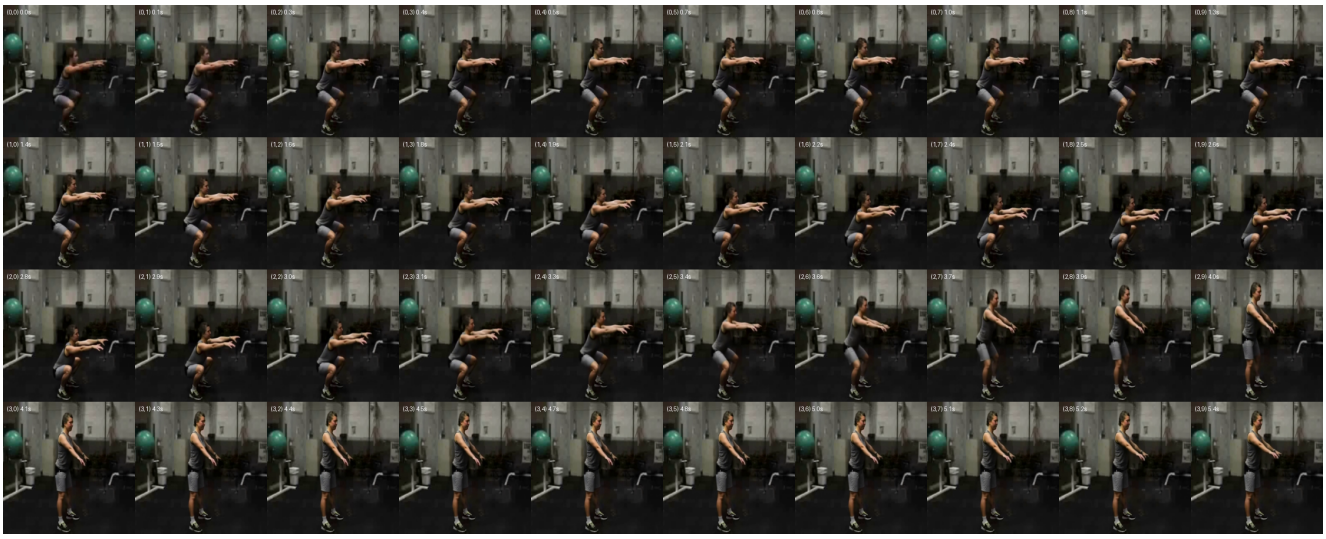

Figure 15. Example of the 4×10 grid-panel layout used to prompt MLLMs. Shown here is a video generated by **Hunyuan** [32] for the action class *BodyWeightSquats*. We uniformly sample 40 frames and place them in row-major order. Each cell overlays its grid coordinates `(row,col)` in the top-left; when the video duration is available, the timestamp is also shown (e.g., `(0,3) 1.2s`). This grid preserves temporal progression and spatial structure, providing clearer visual evidence than direct video input.

G.1. Prompt used for MLLM evaluation

We provide the exact instruction shown to MLLMs for all panel-based evaluations.

```
You are an expert evaluator of AI-generated video quality.

Your job is to analyze a grid of frames extracted from ~5 seconds of video
and score TWO axes:

1. Action Consistency (action_consistency):
   How well does the visible action in the frames match the described target action?
   - Focus only on what is clearly shown.
   - Check pose, motion pattern, timing, and repeated evidence of that action.
   - Do NOT guess intentions outside the frames.

2. Temporal Coherence (temporal_coherence):
   How physically realistic / plausible are the motions and body configurations?
   - Look for broken limbs, impossible joint angles, teleporting limbs,
     limbs merging into objects, obvious gravity violations,
     ghost artifacts (extra arms / missing torso), etc.
   - Minor render glitches are OK if motion is still basically human-plausible.
   - Very warped anatomy or impossible motion should score low.

You must base your judgment ONLY on what is visible in the panel.
Treat the panel as a time grid read left-to-right, top-to-bottom.

Return JSON ONLY with:
- "action_consistency": float in [0,1]
- "temporal_coherence": float in [0,1]
- "confidence": float in [0,1]  (your confidence in these scores)
```

```
- "evidence": a list of AT MOST 3 items. Each item is { "cell": {"row": int,
"col": int}, "description": "short reason" }. Choose the 1-3 most diagnostic cells.
- "rationale": one or two concise sentences summarizing both axes

Scoring guide for BOTH action_consistency and temporal_coherence:
0.90-1.00: very strong evidence, consistent across many frames
0.70-0.89: mostly good, only small issues
0.40-0.69: mixed; frequent issues or uncertainty
0.10-0.39: mostly wrong / implausible / inconsistent
0.00-0.09: completely wrong, impossible, or not supported

Important:
- Use 0-indexed coordinates (row, col) when citing evidence.
- Do NOT invent frames you cannot see.
```

G.2. Limited impact of in-context learning.

We also test whether lightweight in-context learning can better align Qwen3's scores with human judgments. Concretely, we select two *JumpingJack* videos that are not part of the main evaluation set: one clip that clearly satisfies both *Action Consistency* and *Temporal Coherence* (good example) and another clip with severe violations (bad example). These examples, together with short natural-language rationales explaining why they should receive high or low scores, are inserted into the user/assistant prompt as demonstrations before querying the model on the 300 evaluation videos.

Method	Corr. with Action Consistency	Corr. with Temporal Coherence
Qwen3-VL (baseline prompt)	0.34	0.28
Qwen3-VL (in-context learning, JumpingJack demos)	0.29	0.32

Table 11. Correlation (Spearman's ρ) between model predictions and human scores for *Action Consistency* and *Temporal Coherence* on the 300-video evaluation set.

Table 11 reports Spearman's ρ between Qwen3's predictions and human scores. Relative to the baseline prompt, in-context learning *reduces* the correlation for *Action Consistency* from 0.34 to 0.29, while it *increases* the correlation for *Temporal Coherence* from 0.28 to 0.32. Thus, this in-context learning scheme introduces a trade-off between the two dimensions rather than yielding consistent gains, highlighting the need for structured, human-centric evaluation strategies—such as our learned action manifold.

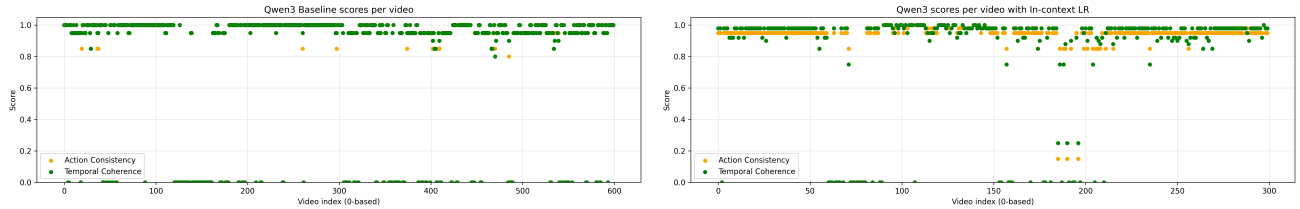

Figure 16. Per-video Qwen3-VL scores for *Action Consistency* (orange) and *Temporal Coherence* (green) before (left) and after (right) applying in-context learning. Individual dots correspond to the model prediction for each generated video. Scores are highly saturated near 0 or 1 in both cases, indicating binary-like decisions rather than nuanced motion reasoning.

Figure 16 visualizes the score distributions. In both baseline and in-context learning settings, predictions are heavily saturated near 0 or 1, indicating that Qwen3-VL often makes binary-like decisions rather than demonstrating nuanced understanding of motion quality. While in-context learning slightly increases the number of mid-range scores for *Temporal Coherence*, the effect is limited and affects only a small subset of videos. This lack of broader distributional shift explains the minimal gain in correlation and suggests that prompting alone cannot adequately capture anatomical or temporal coherence in human motion.