Fréchet Video Motion Distance:
A Metric for Evaluating Motion Consistency in Videos

Jiahe Liu [1 2] **Youran Qu** [3] **Qi Yan** [1 2] **Xiaohui Zeng** [4] **Lele Wang** [1] **Renjie Liao** [1 2 5]

Abstract

Significant advancements have been made in video generative models recently. Unlike image generation, video generation presents greater challenges, requiring not only generating high-quality frames but also ensuring temporal consistency across these frames. Despite the impressive progress, research on metrics for evaluating the quality of generated videos, especially concerning temporal and motion consistency, remains underexplored. To bridge this research gap, we propose *Fréchet Video Motion Distance (FVMD)* metric, which focuses on evaluating motion consistency in video generation. Specifically, we design explicit motion features based on key point tracking, and then measure the similarity between these features via the Fréchet distance. We conduct sensitivity analysis by injecting noise into real videos to verify the effectiveness of FVMD. Further, we carry out a large-scale human study, demonstrating that our metric effectively detects temporal noise and aligns better with human perceptions of generated video quality than existing metrics. Additionally, our motion features can consistently improve the performance of Video Quality Assessment (VQA) models, indicating that our approach is also applicable to unary video quality evaluation. Code is available at https://github.com/ljh0v0/FMD-frechet-motion-distance.

1 Introduction

Recently, diffusion models have demonstrated remarkable capabilities in high-quality image generation (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020). This advancement has been extended to the video domain, giving rise to text-to-video diffusion models (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Zhou et al., 2022; He et al., 2022). Compared to prior works, state-of-the-art diffusion-based video generation models, such as Sora (Brooks et al., 2024), not only aim to generate visually impressive videos but also focus on challenges involving diverse and complex motions, including intricate human dance videos, thrilling fight scenes in movies and sophisticated camera movements. In this case, measuring the motion consistency of these generated videos emerges as a significant research question.

Despite the rapid development of video generation models, research on evaluation metrics for video generation remains insufficient. Currently, FID-VID (Balaji et al., 2019) and FVD (Unterthiner et al., 2018) are widely used to measure the quality of generated videos. FID-VID assesses the visual quality of generated videos by comparing synthesized video frames to real reference video frames, neglecting the video motion quality. In contrast, FVD introduces an evaluation of temporal coherence by extracting video features using a pre-trained action recognition model, Inflated 3D Convnet (I3D) (Carreira & Zisserman, 2017). Recently, VBench provides a comprehensive 16-dimensional evaluation suite for text-to-video generative models (Huang et al., 2023). Nevertheless, the evaluation protocols in VBench for temporal consistency, such as temporal flickering and motion smoothness, tend to award videos with smooth or even static movement, while overlooking high-quality videos with intensive motion, such as dancing and sports videos. Consequently, there is currently no metric specifically designed to evaluate the complex motion patterns in generated videos. This oversight is particularly evident in tasks like motion guided video generation. In these tasks, FID-VID and FVD can only measure whether the appearance of the generated video is consistent with the reference video, but not whether the motion matches the target motion. For VBench, both high-quality and low-quality videos are favored in terms of dynamic degree and penalized for temporal flickering and motion smoothness. This is because the ground-truth videos exhibit intense movements, leading to VBench giving inconsistent assessments compared to human judgement.

To address this research gap, we propose the *Fréchet Video Motion Distance (FVMD)*, a novel metric that focuses on the

Bad (rank 5)　　　　　　　**Human Perception**　　　　　　　**Good** (rank 1)

Model 1　　　　Model 2　　　　Model 3　　　　Model 4　　　　Model 5

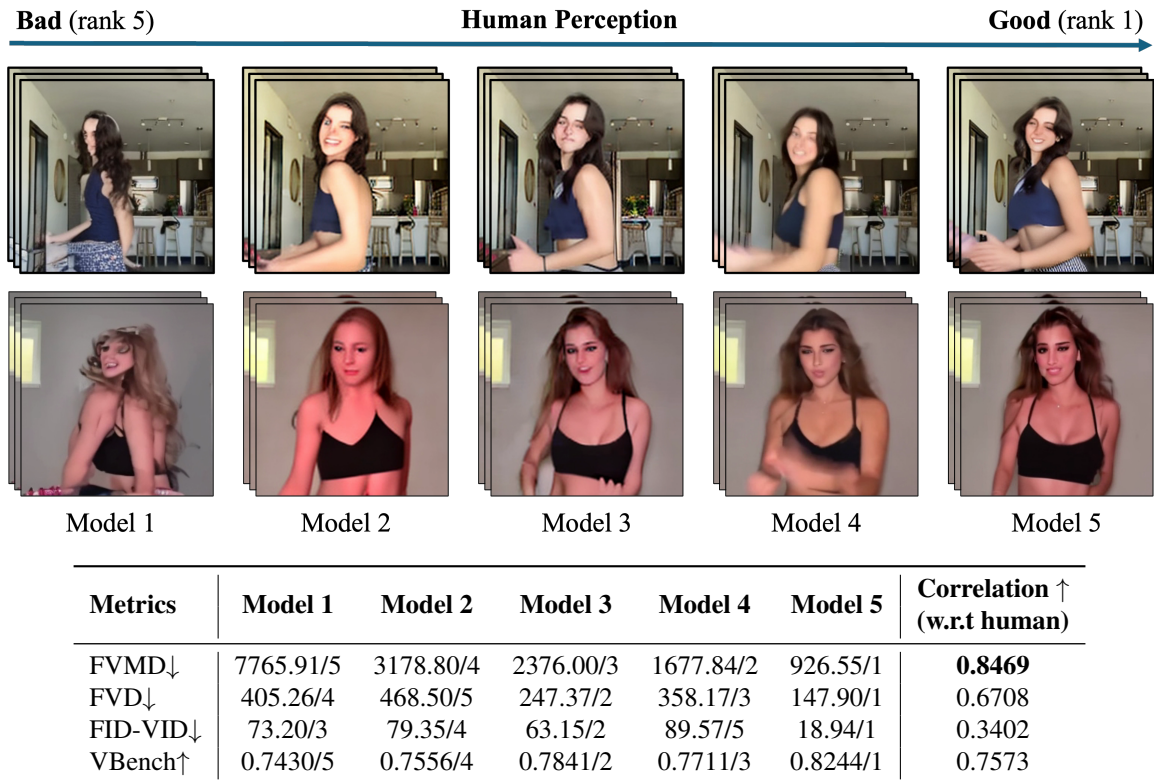| **Metrics**	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Correlation ↑** **(w.r.t human)**
FVMD↓	7765.91/5	3178.80/4	2376.00/3	1677.84/2	926.55/1	**0.8469**
FVD↓	405.26/4	468.50/5	247.37/2	358.17/3	147.90/1	0.6708
FID-VID↓	73.20/3	79.35/4	63.15/2	89.57/5	18.94/1	0.3402
VBench↑	0.7430/5	0.7556/4	0.7841/2	0.7711/3	0.8244/1	0.7573

Figure 1: **Comparison of the fidelity of different video evaluation metrics. Top**: we present videos generated by various models trained on the TikTok dataset (Jafarian & Park, 2022), ranked according to the human ratings in the user study. **Bottom**: we show quantitative scores and relative ranking given by our FVMD and other widely-used metrics, including FVD (Unterthiner et al., 2018), FID-VID (Balaji et al., 2019), and VBench (Huang et al., 2023). The correlations are computed using the Pearson correlation coefficient with human scores (detailed in Section 4.4). Our FVMD achieves the best correlation with human judgment among all the metrics and clearly distinguishes video samples of different quality.

motion consistency of video generation. Our main idea is to measure temporal motion consistency based on the patterns of velocity and acceleration in video movements, as motions conforming to real physical laws should not exhibit sudden changes in acceleration. Specifically, we extract the motion trajectory of key points in videos using a pre-trained point tracking model, PIPs++ (Zheng et al., 2023), and compute the velocity and acceleration for all key points across video frames. We then obtain the motion features based on the statistics of the velocity and acceleration vectors. Finally, we measure the similarity between the motion features of generated videos and ground-truth videos using Fréchet distance (Dowson & Landau, 1982). Our key contributions are as follows: 1) We propose the *Fréchet Video Motion Distance (FVMD)*, a novel metric for video generation focusing on motion consistency. 2) We conduct extensive experiments to evaluate our metric, including sensitivity analysis and human studies, demonstrating our metric is effective in capturing temporal noise and aligns better with human perceptions of video quality than existing metrics. 3) When applied to the Video Quality Assessment (VQA) task, our proposed motion feature leads to consistently improved

performances, suggesting the universality of our method and its potential for generic video evaluation tasks.

2 Related Work

Video Generation. Video generation has long been a challenging and essential area of research. Previous studies have explored various model architectures to tackle this task, such as recurrent neural networks (RNNs) (Babaeizadeh et al., 2017; Castrejon et al., 2019; Denton & Fergus, 2018; Franceschi et al., 2020; Lee et al., 2018), autoregressive transformers (Yan et al., 2021; Wu et al., 2022a; Hong et al., 2022; Ge et al., 2022; Villegas et al., 2022), normalizing flows (Blattmann et al., 2021; Dorkenwald et al., 2021), and generative adversarial networks (GANs) (Vondrick et al., 2016; Saito et al., 2017; Wang et al., 2019; Skorokhodov et al., 2022; Voleti et al., 2022).

Recently, diffusion models have proven to be powerful tools for image generation tasks and have since been applied to the video field, starting with unconditional generation. VDM (Ho et al., 2022b) presents the first results on video generation based on diffusion models by inserting additional

temporal attention blocks into the original 2D U-Net model. Make-A-Video (Singer et al., 2022) and Imagen Video (Ho et al., 2022a) both propose cascaded spatial-temporal up-sampling pipelines to generate long videos with high resolution. LVDM (He et al., 2022) follows the latent diffusion paradigm, lightening and accelerating the video diffusion model by adapting it to the low-dimensional 3D latent space.

Beyond unconditional video generation, significant advancements have been made in video generation conditioned on other modalities, inspired by the success of conditional models like ControlNet in the image domain (Zhang et al., 2023). One notable area is pose-guided video generation, where the goal is to generate videos that adhere to a specified pose sequence, providing control over the motion in the video. Disco (Wang et al., 2023) leverages ControlNet and proposes a novel model architecture with disentangled control to improve the compositionality of human dance synthesis. Animate Anyone (Hu et al., 2023) and Magic Animate (Xu et al., 2023) improve on Disco by adding a motion module to maintain temporal consistency.

Video Evaluation Metrics. Quantitative evaluation metrics can be categorized into frame-level and video-level metrics. Commonly employed frame-level metrics include the Fréchet Inception Distance (FID-VID) (Balaji et al., 2019), Peak Signal-to-Noise Ratio (PSNR) (Wang et al., 2004), Structural Similarity Index Measure (SSIM) (Wang et al., 2004), and CLIP similarity (Radford et al., 2021). FID-VID assesses the generated frames by extracting image features using a pre-trained image classification model, Inception v3 (Szegedy et al., 2016), fitting a Gaussian distribution, and measuring the Fréchet Distance with the ground-truth frames. PSNR is a coefficient representing the ratio between the peak signal and Mean Squared Error (MSE). SSIM is a pixel-level metric that evaluates the luminance, contrast, and structure between generated and reference frames. CLIP similarity measures the alignment between image and text features obtained by the pre-trained CLIP model.

Compared to frame-level metrics, which focus solely on the quality of individual frames, video-level metrics capture both the temporal coherence of a video and the quality of each frame. The Fréchet Video Distance (FVD) (Unterthiner et al., 2018) is a widely used video-level metric. It follows the assumptions of FID and replaces the image classification model with a pre-trained Inflated 3D Convnet (I3D) (Carreira & Zisserman, 2017). Similar to FVD, Kernel Video Distance (KVD) (Unterthiner et al., 2018) employs the same I3D model but utilizes the Maximum Mean Discrepancy (MMD) to measure similarity. Video Inception Score (IS) (Saito et al., 2020) calculates an inception score based on 3D ConvNets (C3D) (Tran et al., 2015).

Recently, VBench (Huang et al., 2023) has been proposed to provide a comprehensive benchmark suite that dissects video quality into hierarchical dimensions, each with tailored prompts and evaluation protocols. The motion-related metrics include temporal flickering, motion smoothness, and dynamic degree. Temporal flickering detects video inconsistency by computing the Mean Absolute Error (MAE) across frames. Motion smoothness evaluates the MAE between the generated frames and synthetic frames using frame interpolation. Since the first two dimensions tend to favor static videos, dynamic degree, which measures the extent of motion in the video, is proposed to counter this effect. However, VBench has considerable limitations, particularly for videos involving intensive motion. For instance, in the task of generating TikTok dancing videos, VBench does not clearly distinguish between high-quality and low-quality samples (see Section 4.4 for details). This is because both types of videos exhibit a high degree of dynamics and large differences between adjacent frames due to the large amplitude of movements. In contrast, even in the presence of intensive motion, our FVMD prefers high-quality videos over low-quality ones, resulting in more accurate scoring.

3 Method

We propose the Fréchet Video Motion Distance (FVMD), a new video generation metric that measures the discrepancy in motion features between generated videos and ground-truth videos. The overall pipeline is illustrated in Figure 2.

3.1 Motion Feature Extraction

Video Key Point Tracking. To construct video motion features, we first track key point trajectories across the video sequence. We utilize the PIPs++ model (Zheng et al., 2023), a state-of-the-art key point tracking approach built upon the particle video method (Sand & Teller, 2008), for this purpose. The selection of PIPs++ is motivated by two key benefits: 1) PIPs++ predicts plausible positions for missing objects in the presence of occlusions, out-of-bounds movements, or difficult lighting conditions. This capability is essential for obtaining a consistent and robust motion trajectory, especially in generated videos where objects may become distorted, blurred, or abruptly vanish. 2) PIPs++ estimates the trajectory of every tracking target independently, allowing computation to be shared between particles within a video, which enhances the speed of inference.

For a set of m generated videos, denoted as $\{X^{(i)}\}_{i=1}^m$, the tracking process starts by truncating longer videos into segments of F frames with an overlap stride of s. Subsequently, we query N target points in a grid shape on the initial frames. PIPs++ is then engaged to estimate N trajectories, denoted as $\hat{Y} \in \mathbb{R}^{F \times N \times 2}$, for these key points, where each trajectory has a coordinate dimension of 2.

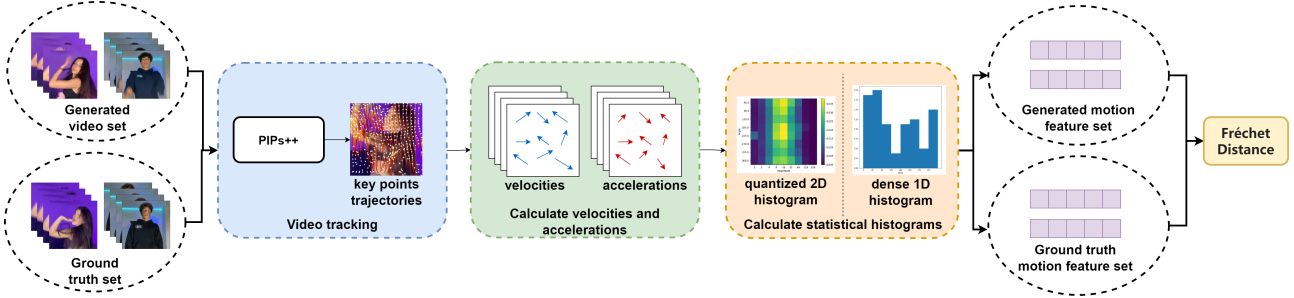**Key Point Velocity and Acceleration Fields.** To obtain

Figure 2: **The overall pipeline of our proposed Fréchet Video Motion Distance (FVMD).** Our pipeline first tracks video key point trajectories using the pre-trained PIPs++ (Zheng et al., 2023) model and computes the velocity and acceleration fields for each frame. The motion features are then derived from the histograms of the quantized velocity and acceleration. FVMD is eventually given by the Fréchet distance between the motion features of generated and ground-truth videos.

representations of motion patterns, we compute the velocity field and acceleration field for each frame within the videos. The temporal inconsistencies in generated videos, such as unnatural changes in object position or posture, sudden deformations or blurring of objects, and jerky movements, can be reflected in disordered key point trajectories, resulting in abrupt changes in the velocity and acceleration of key points. Therefore, the patterns of velocity and acceleration changes over time can effectively indicate whether a video is temporally consistent.

The velocity field $\hat{V} \in \mathbb{R}^{F \times N \times 2}$ measures the first-order difference in key point positions between consecutive frames. To have the same shape as the trajectories $\hat{Y}$, we pad the initial frames in $\hat{V}$ with a zero-frame. The velocity field $\hat{V}$ for a F-frames video segmentation is computed by:

$$\hat{V} = \text{concat}(\mathbf{0}_{N \times 2}, \hat{Y}_{2:F} - \hat{Y}_{1:F-1}) \in \mathbb{R}^{F \times N \times 2}, \quad (1)$$

where $\mathbf{0}_{N \times 2}$ is the zero-padding frame whose subscript indicates its shape. We use $\hat{Y}_{i:j}$ (or $\hat{V}_{i:j}$) to denote the range of frames from the i-th to the j-th inclusively.

Similarly, the acceleration field $\hat{A} \in \mathbb{R}^{F \times N \times 2}$ can be calculated by the first-order difference between the velocity fields in two consecutive frames. Likewise, we pad the first frame of $\hat{A}$ to maintain the same shape as the input:

$$\hat{A} = \text{concat}(\mathbf{0}_{N \times 2}, \hat{V}_{2:F} - \hat{V}_{1:F-1}) \in \mathbb{R}^{F \times N \times 2}, \quad (2)$$

where the subscripts align with those in Equation (1).

Motion Feature. To obtain compact motion features, we further process the velocity and acceleration fields into spatial and temporal statistical histograms. First, we compute the magnitude and angle for each key point in the velocity and acceleration vector fields respectively. Let $\rho(U)$ and $\phi(U)$ denote the magnitude and angle of a vector field U, where $U \in \mathbb{R}^{F \times N \times 2}$ and U can be either $\hat{V}$ or $\hat{A}$. For each frame indexed by $i \in [F]$ and each point indexed by $j \in [N]$, we calculate the magnitude using the l_2 norm and the angle using the inverse hyperbolic tangent $\tanh^{-1}$. The

equations are defined as follows:

$$\rho(U)_{i,j} = \sqrt{U_{i,j,1}^2 + U_{i,j,2}^2}, \forall i \in [F], j \in [N], \quad (3)$$

$$\phi(U)_{i,j} = \left| \tanh^{-1}\left(\frac{U_{i,j,1}}{U_{i,j,2}} \right) \right|, \forall i \in [F], j \in [N]. \quad (4)$$

Next, the magnitudes ρ are clipped to a range of $[0, 255]$. Given that most vector fields have small magnitudes, a base-2 logarithmic transformation is applied for normalization. The magnitudes are then quantized to the nearest integer, resulting in nine discrete bins in range of $[0, 8]$. We also quantize the angle representations ϕ into 8 bins, with each bin encompassing an angle range of 45 degrees.

We employ two methods for calculating statistical histograms of the quantized magnitudes and angles. The first utilizes a quantized **2D histogram**. We divide the F-frame video segments into smaller volumes of size $f \times k \times k$, where f is the number of frames and k the spatial dimensions of each volume. In each volume, we aggregate all vectors to form a 2D histogram, with x and y coordinates corresponding to magnitudes and angles, respectively. These 2D histograms are then concatenated and flattened into a vector, forming the motion feature for the respective video segment. The shape of the quantized 2D histogram is $\frac{F}{f} \times \frac{\sqrt{N}}{k} \times \frac{\sqrt{N}}{k} \times 72$, where the number 72 is derived from 8 discrete bins for angle and 9 bins for magnitude.

Inspired by the HOG (Histogram of Oriented Gradients) approach (Dalal & Triggs, 2005), which counts occurrences of gradient orientation in localized portions of an image, we compile a 2D histogram into a dense **1D histogram** focused on the angle dimension. Similarly, we divide each video segmentation into small $f \times k \times k$ volumes. Our goal is to create a 1D histogram with 8 bins, each corresponding to a range of quantized angles. Within each volume, magnitudes are summed directly into the appropriate angle bin, resulting in an 8-point histogram per volume. By combining these histograms from all volumes, we create the final motion feature, shaped as $\frac{F}{f} \times \frac{\sqrt{N}}{k} \times \frac{\sqrt{N}}{k} \times 8$.
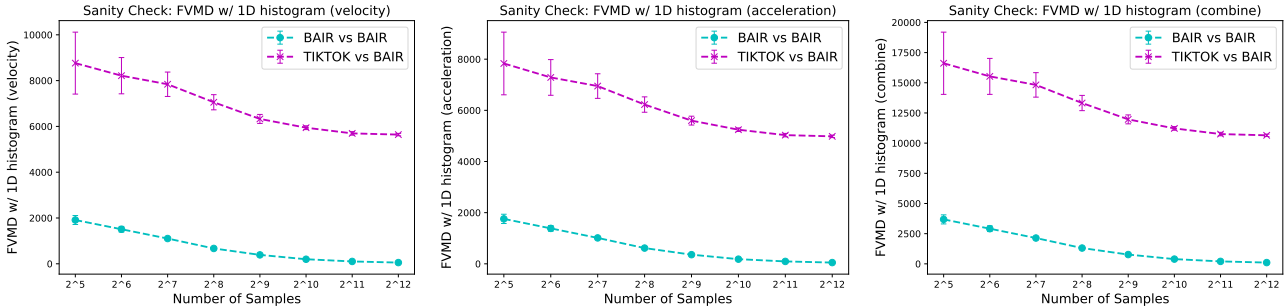
Figure 3: **Sanity check experiments.** We use dense 1D histograms based on velocity, acceleration, and their concatenated combination to construct FVMD metrics. As sample size increases, same-dataset discrepancies (BAIR vs BAIR) converge to zero, while cross-dataset discrepancies (TIKTOK vs BAIR) remain large, verifying the soundness of our FVMD metric.

We apply these two histogram counting methods to separately build the motion features for both velocity and acceleration fields. Additionally, we explore concatenating features from these fields to form a combined motion feature, which can then be used to compute similarity.

3.2 Fréchet Video Motion Distance

After extracting motion features from video segments of generated and ground-truth video sets, we measure their similarity using the Fréchet distance (Dowson & Landau, 1982), which we have named the **Fréchet Video Motion Distance (FVMD)**:

$$d_F(P_{\text{data}}, P_{\text{gen}}) = \left(\inf_{\gamma \in \Gamma(P_{\text{data}}, P_{\text{gen}})} \int \|x - y\|_2^2 d\gamma(x,y)\right)^{\frac{1}{2}}, \quad (5)$$

where the P_{gen} denotes the distribution of motion features for generated videos, P_{data} denotes the distribution of motion features for ground-truth videos, and $\Gamma(P_{\text{data}}, P_{\text{gen}})$ is the set of all couplings of P_{gen} and P_{data}. However, P_{data} and P_{gen} are normally intractable and there is no analytic expression for the Fréchet distance between two arbitrary distributions. Hence, we follow FID (Balaji et al., 2019) to approximate the distributions with multivariate Gaussians. In this case, the Fréchet distance has a closed-form solution:

$$d_F = \|\mu_{\text{data}} - \mu_{\text{gen}}\|_2^2 + \text{tr}\left(\Sigma_{\text{data}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{data}}\Sigma_{\text{gen}})^{\frac{1}{2}}\right), \quad (6)$$

where the μ_{data} and μ_{gen} are the means, and Σ_{data} and Σ_{gen} are the variances. In practice, we use empirical mean and covariance estimations to compute the FVMD.

4 Experiments

In Section 4.2, we conduct a sanity check to verify the soundness of our proposed motion features. In Section 4.3, we conduct sensitivity analysis to demonstrate that our metric is capable of capturing temporal noise. In Section 4.4, we carry out large-scale human studies to show that our FVMD is better aligned with human judgment than the existing metrics. Further, in Section 4.5, we show that our motion features consistently enhance the performance of Video Quality Assessment (VQA) models, suggesting their potential for unary evaluation tasks.

4.1 Implementation Details

We truncate the whole video into segments of $F = 16$ frames using a stride of $s = 1$ and reshape the spatial size of each frame to 256×256. We set the number of tracking points to be $N = 400$ in all experiments. For the motion feature, we set the small volume shape as $4 \times 5 \times 5$ ($f = 4, k = 5$), so that the quantized 2D histogram feature dimension will be $4 \times 4 \times 4 \times 72$. Similarly, the shape of the dense 1D histogram feature is $4 \times 4 \times 4 \times 8$. We empirically identify the velocity-acceleration combined motion feature with a dense 1D histogram as the optimal configuration for our FVMD metric, and thus, it is used as the default.

4.2 Sanity Check

To verify the efficacy of the extracted motion features in representing the motion pattern across a set of videos, we perform a sanity check. We sample two non-overlapping subsets of videos, randomly drawn from the BAIR video pushing dataset (Ebert et al., 2017), with different sample sizes. We then evaluate our metrics on these two subsets. As claimed in the previous work (Unterthiner et al., 2018), the larger the sample size, the better these estimations will be, and the better Fréchet distance reflects the true underlying distance between the distributions. As shown in Figure 3, our metrics converge to zero as the sample size increases, verifying the hypothesis that the underlying motion distribution within the same dataset should remain consistent.

Furthermore, we extract two subsets of equal sample sizes from two distinct datasets, BAIR video pushing and TikTok dancing (Jafarian & Park, 2022). Our FVMD on these two subsets decreases with the increasing sample size, yet remains higher than the FVMD on two subsets within the same dataset. This observation is in accordance with the assumption that the underlying motion distributions of two different datasets should have a larger gap than the ones within the same dataset. Refer to Appendix B.1 for more sanity check results of FVMD with 2D histogram.
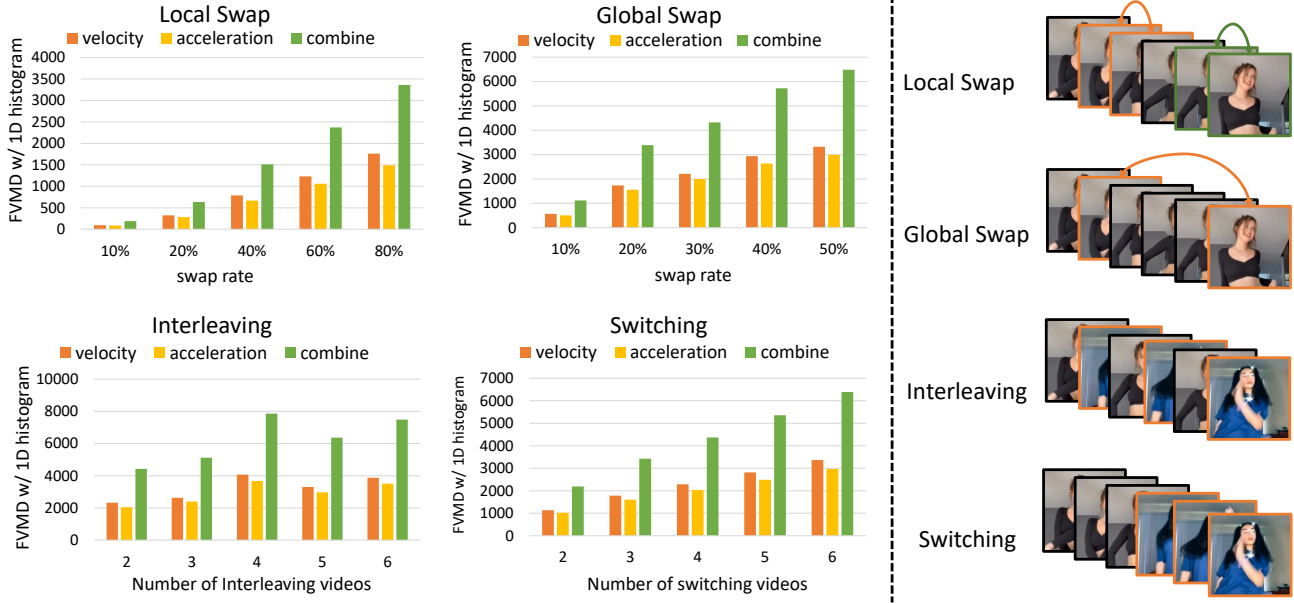
Figure 4: **Sensitivity analysis.** We present the FVMD results in the presence of various temporal noises. FVMD based on combined velocity and acceleration features shows the most reliable performance in distinguishing temporal inconsistencies.

4.3 Sensitivity Analysis

Following the setting of Unterthiner et al., 2018, we validate whether our metrics are sensitive to the temporary inconsistency by adding temporal noise to the TikTok dancing dataset (Jafarian & Park, 2022). We consider four types of temporary noises: 1) local swap: swapping a fraction of consecutive frames in the video sequence, 2) global swap: swapping a fraction of frames in the video sequence with randomly chosen frames, 3) interleaving: weaving the sequence of frames corresponding to multiple different videos to obtain new videos, 4) switching: jumping from one video to another video. As shown in Figure 4, our metrics show a strong capability in differentiating various types of injected noise. In particular, the FVMD based on velocity and acceleration combined features has the best performance. Refer to Appendix B.1 for more results.

4.4 Human Study

An effective video evaluation metric must align with human perceptions. We conduct large-scale human studies to validate our proposed FVMD metric. First, we train a number of conditional video generative models on the TikTok dataset (Jafarian & Park, 2022) and draw video samples from their checkpoints. We then ask users to compare samples from each pair of models to form a ground-truth user score. Subsequently, we calculate the correlation between the score given by each metric and the ground-truth score.

Specifically, we train three different human pose-guided generative models: DisCo (Wang et al., 2023), Animate Anyone (Hu et al., 2023), and Magic Animate (Xu et al.,

2023). We fine-tune these models with distinct architectures and hyper-parameters settings, obtaining over 300 checkpoints with different sample qualities. We evaluate all checkpoints using our FVMD metric and compare the results with FID-VID (Balaji et al., 2019), FVD (Unterthiner et al., 2018), SSIM (Wang et al., 2004), PSNR (Wang et al., 2004), and VBench (Huang et al., 2023), which are the most commonly used video evaluation metrics (Melnik et al., 2024). For VBench, we evaluate five dimensions related to video quality: subject consistency, temporal flickering, motion smoothness, dynamic degree, and imaging quality. Background consistency and aesthetic quality are discarded as they do not support custom videos, and other text prompt-based evaluation dimensions are excluded as inapplicable. We follow the official VBench protocol to calculate an average score for the selected dimensions.

Model Selection. Following the model selection strategy in Unterthiner et al., 2018, we design two settings for the human studies. The first setup is **One Metric Equal**. In this approach, we identify a group of models that have nearly identical scores based on a selected metric. We then investigate whether the other metrics and human raters can effectively differentiate between these models. Based on the results of the i-th metric, we select three groups of model checkpoints corresponding to the quartile points (*i.e.*, top 25%, 50%, and 75%) of its overall distribution, denoted as $\{G_{i,0}, G_{i,1}, G_{i,2}\}$, respectively. Each group contains four models with similar scores for the given metric: $G_{i,j} = \{g_{i,j}^{(0)}, g_{i,j}^{(1)}, g_{i,j}^{(2)}, g_{i,j}^{(3)}\}, j \in \{0, 1, 2\}$. Each model generates a number of videos, forming groups of video sets denoted as $S_{i,j} = \{S_{i,j}^{(0)}, S_{i,j}^{(1)}, S_{i,j}^{(2)}, S_{i,j}^{(3)}\}$. We then create six pairs

Metrics	Eql. FVD	Eql. FID-VID	Eql. SSIM	Eql. PSNR	Eql. VBench-AVG	Eql. FVMD
FVD	-	0.3596	0.0772	-0.1812	-0.1898	-0.7151
FVD-VID	-0.1164	-	0.3061	-0.0944	-0.5226	-0.6956
SSIM	-0.5926	-0.7853	-	-0.8130	-0.7973	0.0527
PSNR	-0.4267	-0.6096	0.1204	-	-0.7973	0.3031
VBench-subject consistency	-0.2823	-0.0386	0.262	-0.2135	-0.601	-0.8691
VBench-temporal flickering	-0.5803	-0.2081	0.2413	-0.128	-0.4198	-0.5938
VBench-motion smoothness	-0.4708	-0.1561	0.2698	-0.0818	-0.3844	-0.5448
VBench-dynamic degree	0.5411	0.238	-0.4063	0.0338	0.2823	0.3894
VBench-imaging quality	0.5684	0.4093	0.7160	0.8859	0.7062	0.0383
VBench-AVG	0.5112	0.8835	0.341	0.2097	-	-0.5769
FVMD	0.9170	**0.9184**	**0.7191**	**0.4790**	**0.7348**	-
Combine FVMD & FVD	**0.9173**	0.8441	0.2886	0.0383	0.4860	-
Agreement rate (human raters)	0.6773	0.8196	0.7653	0.7184	0.7980	0.7461

Table 1: **Pearson correlation for One Metric Equal experiments.** This table shows the Pearson correlation between metrics scores and human perceptions when one selected metric is almost equal, *i.e.*, one can not distinguish these videos relying on the given metric alone. The correlation ranges from 1 to −1, with values closer to 1 (−1) indicating stronger positive (negative) correlation. We also report the agreement rate among raters as a percentage from 0 to 1. Overall, our FVMD demonstrates the strongest capability to distinguish videos when the other metrics fall short.

of video sets from any two video sets[1] for human studies.

The second setting, **One Metric Diverse**, evaluates the agreement among different metrics when a single metric provides a clear ranking of the performances of the considered video generative models. Specifically, we select model checkpoints whose samples can be clearly differentiated according to the given metric and test the consistency between this metric and other metrics as well as human raters. Similar to the above setups, we select three groups of models, each comprising four checkpoints with significantly different scores for the given metric. We then draw video samples and construct pairs among them for human studies.

Human Rating. We ask over 200 individuals to evaluate videos produced by the selected models to study how well the evaluation metrics align with human judgment. For every video set pair, we randomly extract three generated videos. Raters are asked to rate all three video pairs across all model pairs and the most frequently selected option is recorded as the final decision. Following this, we aggregate and determine the user scores for each group by calculating the Borda count (Borda, 1781) across all user answers. For more implementation details, please refer to Appendix A.1.

Evaluation Metrics. For each group, we compute the Pearson correlation coefficient between raw scores given by different metrics and the ground-truth human score. Subsequently, the average value across the three groups is computed to represent the final correlation between the metrics and human scores. The higher the value, the better the metric aligns with human judgment.

Results. We compare FVMD based on combined velocity-acceleration features and dense 1D histograms with existing metrics, as shown in Table 1 and Table 2. Additionally, we explore combining the FVMD with FVD using the F1 score. Evidently, our FVMD shows consistently positive and significantly higher correlation coefficients than the other metrics in both the **One Metric Equal** and **One Metric Diverse** settings. The quantitative results imply that FVMD is more trustworthy than the baseline metrics. For more ablation results, please refer to Appendix B.2. We also report the agreement rate among human raters, which is calculated as the fraction of answers consistent with aggregated answer. The high agreement among raters indicates their confidence in the survey, enhancing the human study credibility.

In general, the experimental results indicate that our FVMD aligns more closely with human perception across nearly all experimental settings. In the experiments of **One Metric Equal**, we observe that FVMD significantly outperforms the other metrics, suggesting that in scenarios where the other metrics fail to evaluate video quality, FVMD can serve as an effective metric to help distinguish videos. On the other hand, from the equivalent FVMD column, it is evident that no other metrics can reliably distinguish between models when FVMD results are equal. Moreover, in the experiments of **One Metric Diverse**, FVMD demonstrates generally higher Pearson correlation than the other metrics. Despite some dimensions of VBench aligning more closely with human perception in certain settings, the overall average score provided by VBench still does not surpass FVMD. Therefore, FVMD is more capable of providing a comprehensive assessment of video quality compared to VBench.

---

[1]The pairs are $(S_{i,j}^{(0)}, S_{i,j}^{(1)})$, $(S_{i,j}^{(0)}, S_{i,j}^{(2)})$, $(S_{i,j}^{(0)}, S_{i,j}^{(3)})$, $(S_{i,j}^{(1)}, S_{i,j}^{(2)})$, $(S_{i,j}^{(1)}, S_{i,j}^{(3)})$, and $(S_{i,j}^{(2)}, S_{i,j}^{(3)})$.

Metrics	Diverse FVD	Diverse FID-VID	Diverse SSIM	Diverse PSNR	Diverse VBench-AVG	Diverse FVMD
FVD	0.1007	0.1952	0.2149	0.5662	-0.1935	0.0561
FVD-VID	-0.2080	-0.2002	0.0201	0.3987	-0.4441	-0.0268
SSIM	-0.8617	-0.5556	-0.7600	-0.5515	-0.6404	-0.6832
PSNR	-0.6764	-0.7377	-0.6812	-0.6538	-0.5326	-0.5842
VBench-subject consistency	-0.0102	-0.3691	- 0.0452	0.1914	-0.2321	0.1819
VBench-temporal flickering	-0.5898	0.0755	-0.0870	0.5233	-0.7701	-0.5315
VBench-motion smoothness	-0.4563	0.1822	0.1276	0.5936	-0.6547	-0.2125
VBench-dynamic degree	**0.8285**	-0.3992	0.2223	-0.5731	0.6866	0.7047
VBench-imaging quality	0.5064	0.4593	**0.8505**	0.3655	0.6657	**0.7404**
VBench-AVG	0.7163	0.1720	0.5694	0.4479	0.3031	0.4688
FVMD	0.7321	**0.8561**	0.6921	**0.9677**	**0.7928**	0.6808
Combine FVMD & FVD	0.5940	0.5621	0.5624	0.8192	0.5245	0.4901
Agreement rate (human raters)	0.8282	0.7336	0.7529	0.7836	0.8132	0.7665

Table 2: **Pearson correlation for One Metric Diverse experiments.** This table shows the Pearson correlation between metrics scores and human perceptions when one metric is diverse, *i.e.*, one can distinguish these videos relying on the give metric alone. We also report the agreement rate among raters, which is a percentage ranging from 0 to 1.

Method	Vanilla		Ours	
	PLCC ↑	SROCC ↑	PLCC ↑	SROCC ↑
VSFA	0.765	0.762	0.779	0.770
FastVQA	0.834	0.832	0.841	0.838
SimpleVQA	0.847	0.840	0.870	0.861

Table 3: **Unary video quality assessment.** Our motion features consistently boost VQA method performance.

Stage	Avg. runtime (sec. per video)
Video tracking	1.220
Compute vector fields	0.060
Build 1D histogram	0.018
Compute Fréchet distance	0.002
Overall	1.325

Table 4: **Inference time.** Most of the runtime is due to video tracking, while other components are light in computation.

video tracking stage due to the PIPs++ model.

4.5 Unary Evaluation

FVMD is a pair-wise metric that provides a robust assessment score when a ground-truth video set is available. However, when access to a ground-truth video set is not possible, unary video quality assessment methods become necessary (Liu et al., 2024). Therefore, we extend the application of our explicit motion features to the Video Quality Assessment (VQA) tasks. We adapt open-source state-of-the-art VQA backbones, including SimpleVQA (Sun et al., 2022), FastVQA (Wu et al., 2022b) and VSFA (Li et al., 2019), to incorporate our motion feature. We compare their empirical performance on the KVQ dataset (Lu et al., 2024), which is a large-scale VQA benchmark dataset with over 4,000 user-created video clips. We compare our predicted Mean Opinion Score (MOS) score with the ground-truth score using Pearson linear correlation coefficient (PLCC) and Spearman rank-order correlation coefficients (SROCC). The results are shown in Table 3. The performances of VQA models are clearly enhanced by our explicit motion features.

4.6 Efficiency

We report the inference time for each stage of the FVMD pipeline in Table 4. We test the runtime on two subsets of the Tiktok dataset consisting of 1024 16-frame 256×256 videos. The majority of the runtime is consumed by the

5 Conclusion

In this work, we propose a novel metric, *Fréchet Video Motion Distance (FVMD)*, to evaluate sample quality for video generative models with a focus on temporal coherence. We design an explicit motion representation based on the patterns of velocity and acceleration in video movements. Our metric compares the discrepancies of these motion features between the generated and ground-truth video sets, measured by the Fréchet distance. We conduct both sensitivity analysis and human studies to evaluate the effectiveness of our proposed metric. Our proposed FVMD outperforms existing metrics in many aspects, such as better alignment with human judgment and a stronger capability to distinguish videos of different quality. Moreover, we validate the promising potential of our motion features for unary video quality assessment through experiments on VQA tasks.

For future directions, we aim to explore a more comprehensive motion representation that conforms to the physical laws of object movement in the real world. This will help detect physically implausible motions and interactions in AI-generated videos, such as abnormal human movements or object trajectories that defy common sense.

Acknowledgements

References

Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R. H., and Levine, S. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017.

Balaji, Y., Min, M. R., Bai, B., Chellappa, R., and Graf, H. P. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, pp. 2, 2019.

Blattmann, A., Milbich, T., Dorkenwald, M., and Ommer, B. ipoke: Poking a still image for controlled stochastic video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14707–14717, 2021.

Borda, J. d. M'emoire sur les' elections au scrutin. *Histoire de l'Acad'emie Royale des Sciences*, 1781.

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., and Ramesh, A. Video generation models as world simulators. *preprint*, 2024.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Castrejon, L., Ballas, N., and Courville, A. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7608–7617, 2019.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pp. 886–893. Ieee, 2005.

Denton, E. and Fergus, R. Stochastic video generation with a learned prior. In *International conference on machine learning*, pp. 1174–1183. PMLR, 2018.

Dorkenwald, M., Milbich, T., Blattmann, A., Rombach, R., Derpanis, K. G., and Ommer, B. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3742–3753, 2021.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Ebert, F., Finn, C., Lee, A. X., and Levine, S. Self-supervised visual planning with temporal skip connections. *CoRL*, 12:16, 2017.

Franceschi, J.-Y., Delasalles, E., Chen, M., Lamprier, S., and Gallinari, P. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pp. 3233–3246. PMLR, 2020.

Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., and Parikh, D. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022.

He, Y., Yang, T., Zhang, Y., Shan, Y., and Chen, Q. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., and Bo, L. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al. Vbench: Comprehensive benchmark suite for video generative models. *arXiv preprint arXiv:2311.17982*, 2023.

Jafarian, Y. and Park, H. S. Self-supervised 3d representation learning of dressed humans from social media videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018.

Li, D., Jiang, T., and Jiang, M. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 2351–2359, 2019.

Liu, X., Min, X., Zhai, G., Li, C., Kou, T., Sun, W., Wu, H., Gao, Y., Cao, Y., Zhang, Z., et al. Ntire 2024 quality assessment of ai-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6337–6362, 2024.

Lu, Y., Li, X., Pei, Y., Yuan, K., Xie, Q., Qu, Y., Sun, M., Zhou, C., and Chen, Z. Kvq: Kaleidoscope video quality assessment for short-form videos. *arXiv preprint arXiv:2402.07220*, 2024.

Melnik, A., Ljubljanac, M., Lu, C., Yan, Q., Ren, W., and Ritter, H. Video diffusion models: A survey. *arXiv preprint arXiv:2405.03150*, 2024.

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*, pp. 8748–8763. PMLR, 2021.

Saito, M., Matsumoto, E., and Saito, S. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pp. 2830–2839, 2017.

Saito, M., Saito, S., Koyama, M., and Kobayashi, S. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *International Journal of Computer Vision*, 128(10):2586–2606, 2020.

Sand, P. and Teller, S. Particle video: Long-range motion estimation using point trajectories. *International journal of computer vision*, 80:72–91, 2008.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Skorokhodov, I., Tulyakov, S., and Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3626–3636, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Sun, W., Min, X., Lu, W., and Zhai, G. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 856–865, 2022.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

Voleti, V., Jolicoeur-Martineau, A., and Pal, C. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

Vondrick, C., Pirsiavash, H., and Torralba, A. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

Wang, T., Li, L., Lin, K., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. Disco: Disentangled control for referring human dance generation in real world. *arXiv e-prints*, pp. arXiv–2307, 2023.

Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., and Catanzaro, B. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pp. 720–736. Springer, 2022a.

Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., and Lin, W. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pp. 538–554. Springer, 2022b.

Xu, Z., Zhang, J., Liew, J. H., Yan, H., Liu, J.-W., Zhang, C., Feng, J., and Shou, M. Z. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.

Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Zheng, Y., Harley, A. W., Shen, B., Wetzstein, G., and Guibas, L. J. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19855–19865, 2023.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., and Feng, J. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

A More Implementation Details

A.1 Human study

For each model group $G_{i,j} = \{g_{i,j}^{(0)}, g_{i,j}^{(1)}, g_{i,j}^{(2)}, g_{i,j}^{(3)}\}$ and its corresponding video sets group $S_{i,j} = \{S_{i,j}^{(0)}, S_{i,j}^{(1)}, S_{i,j}^{(2)}, S_{i,j}^{(3)}\}$, we ask the rater to compare all six generated pairs. For example, the pair $(S_{i,j}^{(0)}, S_{i,j}^{(1)})$, where $S_{i,j}^{(0)}$ and $S_{i,j}^{(1)}$ are sets of videos generated by model $g_{i,j}^{(0)}$ and model $g_{i,j}^{(1)}$ respectively, we randomly select three video pairs that have the same content from them. The rater needs to compare all these three video pairs. If the rater chooses the video generated by model $g_{i,j}^{(0)}$ for two or more of the pairs, then we consider that the rater prefers model $g_{i,j}^{(0)}$. In this case, model $g_{i,j}^{(0)}$ score 1, and model $g_{i,j}^{(1)}$ score 0.

When all raters have completed scoring the six video set pairs, we will sum the scores obtained by models $\{g_{i,j}^{(0)}, g_{i,j}^{(1)}, g_{i,j}^{(2)}, g_{i,j}^{(3)}\}$ respectively to determine the final user score for each model and rank them accordingly

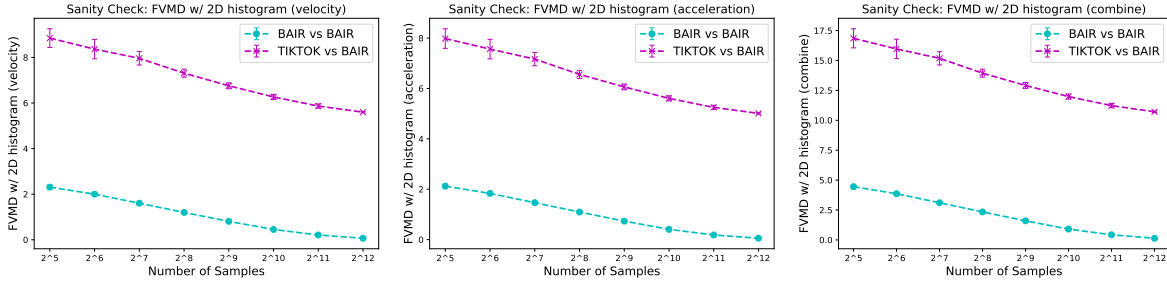# B Addition Experiments Results

B.1 FVMD with 2D histogram

Figure 5: **Sanity check.** We visualize the curve for FVMD with quantized 2D histogram versus the number of samples.

Noise type	Hyperparameter	Int. 1	Int. 2	Int. 3	Int. 4	Int. 5
Local Swap	The proportion of frames swapped	10%	20%	40%	60%	80%
Global Swap	The proportion of frames swapped	10%	20%	30%	40%	50%
Interleaving	The number of videos interleaved	2	3	4	5	6
Switching	The number of videos switched	2	3	4	5	6

Table 5: **Hyperparameter design of the noise study.**

Sanity Check. Similar to the sanity check experiments conducted for FVMD with a 1D dense histogram, we also performed a sanity check for the 2D histogram setting. The results, shown in Figure 5, demonstrate that the 2D histogram feature also supports our hypothesis: the underlying motion distribution within the same dataset remains consistent, while the distribution between two different datasets exhibits a larger gap.

Sensitivity Analysis. We conduct the sensitivity analysis on a subset with a fixed 1024 video clips of the TikTok dataset (Jafarian & Park, 2022). In our sensitivity analysis, the hyperparameter design for the intensity of different types of noise is as shown in Table 5. Figure 6 illustrates the behavior of the FVMD with a 2D histogram when various types of static noise are added to the temporal dimension of videos. It is observable that the values of all implementations of FVMD increase with the escalation of added noise.
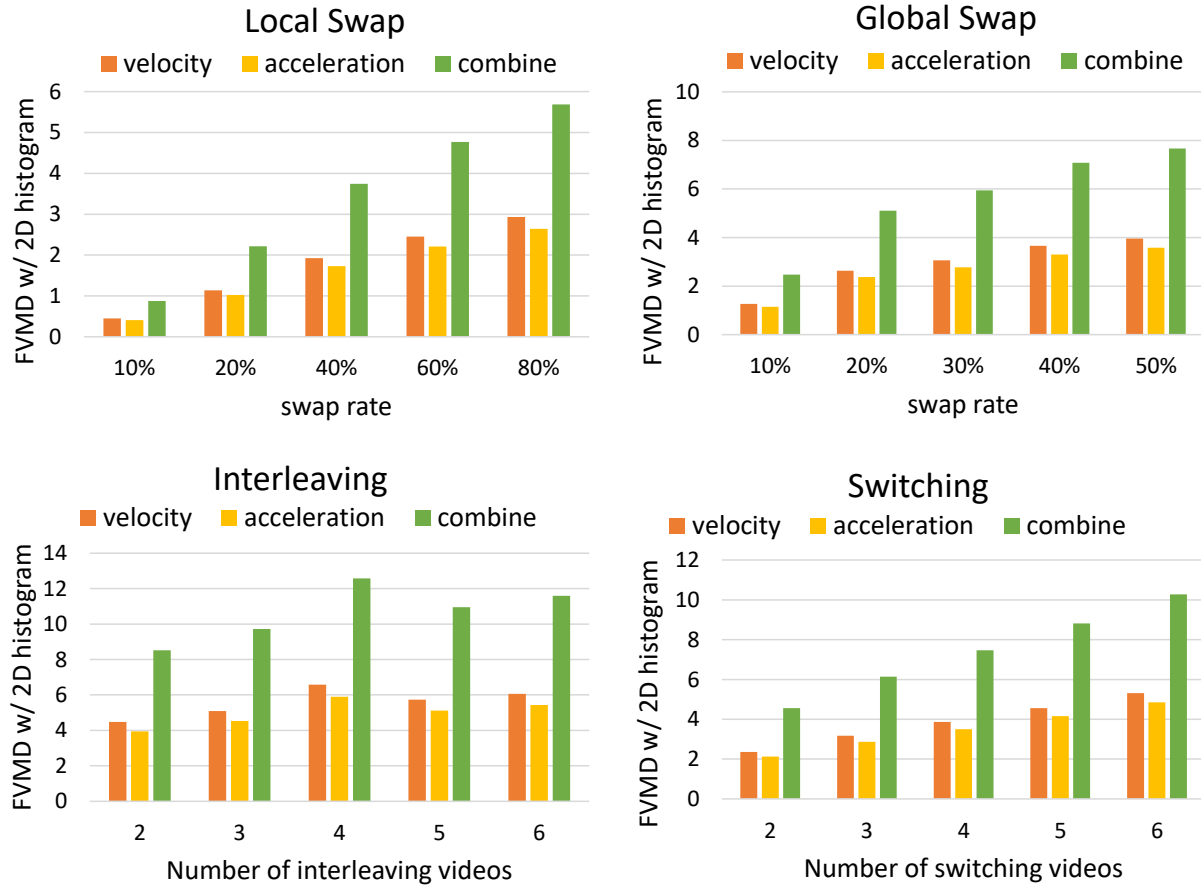
Figure 6: **Sensitivity analysis**. Behaviors of FVMD with 2D histogram when adding various types of static noise to the temporal dimension of videos.

Configuration			eql. FVD	eql. FID-VID	eql. SSIM	eql. PSNR	eql. VBench-overall score
motion representation	statistics	stride s					
velocity	2D	16	0.9314	0.8469	0.4263	0.0448	0.0117
acceleration	2D	16	**0.9316**	0.8453	0.4216	0.0362	-0.0063
combine	2D	16	0.9315	0.8464	0.4241	0.0423	0.0039
velocity	1D	16	0.8773	0.9210	0.4934	0.2007	0.5642
acceleration	1D	16	0.8601	0.8985	0.5104	0.1993	0.5510
combine	1D	16	0.8612	0.9091	0.4925	0.1894	0.5414
velocity	2D	1	0.8555	0.8106	0.3903	0.0080	-0.7144
acceleration	2D	1	0.8627	0.8136	0.3913	0.0115	-0.1101
combine	2D	1	0.8569	0.8115	0.3916	0.0109	-0.1144
velocity	1D	1	0.9172	**0.9253**	0.7128	**0.4920**	**0.7359**
acceleration	1D	1	0.9276	0.9112	0.7162	0.4851	0.7354
combine	1D	1	0.9170	0.9184	**0.7191**	0.479	0.7348

Table 6: **Ablation study on One Metric Equal setting.** The experimental setup is consistent with that described in Table 1. The eql. FMD column has been omitted.

Configuration			divers. FVD	divers. FID-VID	divers. SSIM	divers. PSNR	divers. VBench-overall score	divers. FVMD
motion representation	statistics	stride s						
velocity	2D	16	0.5851	0.2831	0.6209	0.6977	0.3275	0.3985
acceleration	2D	16	0.5791	0.2893	0.6136	0.7013	0.3169	0.3936
combine	2D	16	0.5838	0.2880	0.6189	0.7008	0.3232	0.3963
velocity	1D	16	0.6365	0.6920	0.6126	0.8866	0.6835	0.5768
acceleration	1D	16	0.6282	0.7016	0.6100	0.8929	0.6854	0.5750
combine	1D	16	0.6269	0.6910	0.6085	0.8866	0.6781	0.5699
velocity	2D	1	0.5075	0.2714	0.4803	0.7162	0.3358	0.3942
acceleration	2D	1	0.5076	0.2776	0.4820	0.7196	0.3424	0.3946
combine	2D	1	0.5082	0.2772	0.4823	0.7193	0.3399	0.3945
velocity	1D	1	**0.7388**	**0.8588**	**0.6959**	**0.9685**	0.7951	**0.6836**
acceleration	1D	1	0.7311	0.8582	0.6905	0.9665	**0.7952**	0.6835
combine	1D	1	0.7321	0.8561	0.6921	0.9677	0.7928	0.6808

Table 7: **Ablation study on One Metric Diverse setting** The experimental setup is consistent with that described in Table 2.

B.2 Ablation Study

To determine the optimal configuration for our FVMD, we conduct ablation experiments under the same experimental setup as used in the human study. We explore alternative designs for the motion features, including: 1) different motion representations, including computing only velocity fields, only acceleration fields, and combining velocity and acceleration; 2) different methods for statistically characterizing vector fields, including quantized 2D histograms and dense 1D histograms; 3) the degree of overlap when extracting 16-frame segments from the entire video, ranging from no overlap (stride=16) to maximum overlap (stride=1). The results of the ablation study are presented in Table 6 and Table 7.

It is evident that different motion representations do not significantly impact the performance of our metric. Additionally, the performance of FVMD with a dense 1D histogram surpasses that of FVMD with a quantized 2D histogram. For FVMD with a dense 1D histogram, the maximum overlap strategy when extracting video clips outperforms the non-overlap strategy across all experimental setups. Overall, FVMD utilizing a combined motion representation with a dense 1D histogram and maximum overlap video segments aligns more closely with human perception. Therefore, we select this as the default configuration for our FVMD metric.