

# Beyond the Golden Data: Resolving the Motion-Vision Quality Dilemma via Timestep Selective Training

## Supplementary Material

### 6. Timestep Distribution Analysis

Fig. 5 visualizes the timestep distributions across samples with different MQ and VQ scores under two  $\kappa_{\text{base}}$  configurations. We set  $\kappa_{\text{base}} = 2$  and 4 such that when  $\text{MQ} = \text{VQ}$ , the distributions degenerate to uniform and logit-normal, respectively. Notably, samples with higher MQ scores shift the distribution toward larger timesteps, while those with higher VQ scores concentrate sampling at lower timesteps. This adaptive reweighting allows our approach to dynamically allocate timesteps based on sample-specific quality metrics. However, samples with uniformly low or high MQ and VQ scores exhibit similar relative quality metrics, rendering timestep reweighting alone insufficient for differentiation. To complement this, we incorporate a probabilistic retention strategy during data loading, where each sample is kept with probability  $\max(\text{vq}_{\text{norm}}, \text{mq}_{\text{norm}})$  (see Sec. 3.3).

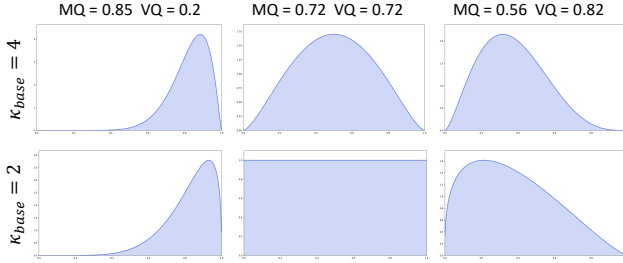


Figure 5. Timestep sampling distributions under varying quality scores. We visualize the Beta distributions with  $\kappa_{\text{base}} = 4$  (top) and  $\kappa_{\text{base}} = 2$  (bottom) for samples with different MQ and VQ scores. Left: High MQ, low VQ shifts toward large timesteps. Middle: Equal MQ and VQ yields centered distributions. Right: Low MQ, high VQ concentrates on small timesteps. Note that  $\kappa_{\text{base}} = 2$  degenerates to uniform when  $\text{MQ} = \text{VQ}$  (middle-bottom).

### 7. Computational Analysis

Our proposed training strategy requires VQ and MQ scores for each sample, which can be pre-computed offline before training begins. During training, we only perform two lightweight operations: sample dropout and timestep distribution computation based on VQ and MQ scores. The former seamlessly integrates with the dataloader’s prefetch mechanism, while the latter incurs negligible computational overhead. Consequently, our method achieves nearly identical GPU training time compared to the baseline, making it practically cost-free in terms of computational resources.

### 8. Dataset Settings

#### 8.1. Data Distribution Visualization

Fig. 6 illustrates the distribution of MQ and VQ scores across our 280K training samples, scored using VideoAlign [26]. The MQ distribution (left) exhibits clear bimodality with peaks at  $\approx 2.3$  and  $\approx 2.8$ , reflecting the natural divide between static and dynamic video content. In contrast, VQ distribution (right) follows a unimodal, right-skewed pattern centered at  $\approx 2.7$ , with most samples concentrated between 2.0–3.5. We partition this data into four subsets based on median-aligned thresholds ( $\text{MQ}=2.5$ ,  $\text{VQ}=2.7$ ): HMHV (100K), HMLV (80K), LMHV (80K), and LMLV (20K). Note that these distributions reflect our curated dataset where we deliberately selected different amounts from each quality category and mixed them together to ensure balanced training data.

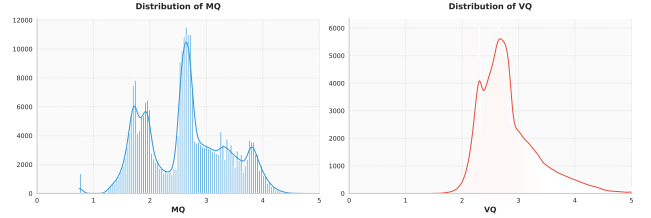


Figure 6. Data distribution of our collected training set.

#### 8.2. Video Caption Generation

To obtain precise and detailed video descriptions, we employ Gemini 2.5 Pro with carefully designed prompts that emphasize comprehensive scene analysis. Our prompting strategy instructs the model to provide structured captions covering eight key dimensions: (1) brief summary, (2) subject details including appearance and attire, (3) environment and context, (4) chronological actions and interactions, (5) aesthetic style, (6) technical visual attributes, (7) cinematography, and (8) camera movement. This structured approach ensures consistent, high-fidelity annotations that capture both visual content and stylistic nuances essential for training high-quality video generation models. Figure 7 illustrates our complete prompting template alongside a representative caption example generated for a wedding celebration scene.



You are an expert video analysis expert specializing in generating dense, high-fidelity captions for video datasets. Your goal is to describe the input video with extreme precision, focusing on visual details, action, and cinematographic attributes. Analyze the provided video and generate a caption that strictly follows the structure and style defined below. Your response must be a continuous text (or comprised of detailed segments) covering these specific aspects in order:

1. **Brief Summary:** A single sentence overview of the main subject and action.
2. **Subject Details:** Detailed description of the main character(s). Include age, ethnicity, hair, specific clothing details (colors, patterns, patches, layers), and initial pose/expression.
3. **Environment & Context:** Describe the background (wall colors, objects, furniture), lighting conditions (bright, dim, natural, artificial), and any inference about time of day or weather (or state if it's indeterminable).
5. **Chronological Action:** Describe the sequence of events. What is the subject doing? Who enters the frame? Describe movements from left to right or background to foreground. Focus on interactions and micro-expressions.
6. **Aesthetic Style:** Describe the visual style (e.g., realistic, cinematic, home video, TV series look). Mention the color palette (muted, vibrant, pastel).
7. **Technical Visuals:** Analyze the image quality terms: saturation (low/medium/high), contrast, lightness, and color tone (neutral, warm, cool).
8. **Cinematography:** Describe the movement of the camera.

# Style Guidelines

**Tone:** Objective, descriptive, and professional.

**Detail Level** Extremely high. Do not just say "a jacket," say "a black and white jacket with red accents and a patch on the left side."

**Format:** Do not use bullet points. Write in full, descriptive sentences.

A blonde woman in a white dress and a man in a patterned shirt embrace and congratulate a central male figure in a white t-shirt, culminating in the woman kissing him on the cheek during a celebratory event. The central subject is a young adult Caucasian male with short, light brown hair, wearing a plain white crew-neck t-shirt; his expression shifts from hearty laughter to a warm, appreciative smile. To his right stands a young adult Caucasian woman with blonde hair styled in a loose updo with strands framing her face, wearing long, dangling crystal earrings and a white dress with spaghetti straps and subtle beaded embellishments. To the left, another male subject appears, wearing a white short-sleeved button-down shirt featuring a dense blue floral or paisley pattern; he has short dark hair and is viewed primarily from the back and side. The scene takes place indoors against a textured, rustic wall that appears to be concrete or plaster with warm, earth-tone hues. In the blurred background, a figure sits behind a table with a laptop featuring a glowing logo, suggesting a DJ booth or reception setup. The lighting is artificial and atmospheric, with a strong backlight creating a halo effect on the woman's hair and a warm, ambient glow throughout the space.

The sequence begins with the central man laughing with his eyes nearly closed as he is embraced. The man in the patterned shirt enters fully from the left, wrapping his arms around the central man in a tight hug, turning his back to the camera. The woman stands close on the right, smiling broadly and looking affectionately at the central man, her hand resting near him. As the hug between the two men continues, the central man turns his head slightly toward the woman, locking eyes with her. The interaction concludes as the woman leans in and presses a kiss onto the central man's left cheek, causing him to smile widely while looking forward. The aesthetic style is that of high-end wedding videography, characterized by a candid, documentary feel that captures genuine emotion. The color palette is warm and inviting, dominated by creamy whites, skin tones, and the amber hues of the background lighting.

Visually, the footage exhibits high saturation, emphasizing the warmth of the scene, with a medium contrast that preserves details in the bright whites of the clothing and the darker shadows of the embrace. The lighting creates a soft, diffused look with a distinct warm color tone, enhancing the romantic and celebratory mood. The cinematography utilizes a static, medium close-up shot with a shallow depth of field, keeping the three subjects in sharp focus while blurring the background to isolate the intimate interaction. The camera remains steady, acting as an observer to the spontaneous affection without distracting movement.

Figure 7. Dense video captioning example. We leverage Gemini 2.5 Pro with structured prompts (top) to generate comprehensive video descriptions. The bottom panel shows an example caption (300+ words) for a wedding scene, demonstrating fine-grained descriptions of subjects' clothing, spatial dynamics, lighting, and camera techniques. This structured approach ensures annotation quality and consistency across our 280K training dataset.

## 9. More Qualitative Comparisons

Additional qualitative video comparisons are included in the supplementary materials, which further demonstrate the effectiveness of our approach.

## 10. Robustness to Scorer Noise

TQD employs probabilistic Beta-distribution sampling rather than hard-threshold filtering, providing inherent resilience to scorer noise—minor scoring errors only cause slight shifts in the sampling center ( $\mu$ ) without excluding any data. To quantify this, we inject 10% and 20% Gaussian noise into MQ and VQ scores on Wan Set-A.

As shown in Table 5, TQD with 20% noise still significantly outperforms the baseline across all metrics. This robustness is further validated by consistent improvements on the independent VBench metrics.

Table 5. Robustness to scorer noise on Wan Set-A. TQD maintains substantial gains over the baseline even with 20% noise.

Method	IQ	Aesthetic	Dynamic	Smooth	MQ	VQ
Uniform (no TQD)	0.6916	0.5722	0.5312	0.9935	2.1388	3.2537
TQD (0% noise)	<b>0.7010</b>	<b>0.5757</b>	<b>0.6384</b>	0.9923	<b>2.2557</b>	<b>3.3450</b>
TQD (10% noise)	0.6985	0.5743	0.6251	0.9931	2.2431	3.3285
TQD (20% noise)	0.6953	0.5731	0.6012	0.9928	2.2114	3.3085