DiVid: Diagnosing Dimension-Specific Diversity Collapse in Video Generation Models

Anonymous authors
Paper under double-blind review

Abstract

Despite remarkable progress, video generation models often produce highly similar outputs when repeatedly sampled from the same prompt, limiting their usefulness for creative exploration. Existing diversity evaluations primarily rely on global scalar metrics, which obscure where diversity collapses in the spatiotemporal space of videos. We introduce **DiVid**, a dimension-level diagnostic framework that decomposes video generation diversity into six interpretable dimensions: Semantic, Style, Subject, Scene, Motion, and Camera. Each dimension is measured through a reproducible computer-vision pipeline and analyzed alongside quality and instruction faithfulness to examine potential trade-offs. Systematic evaluation of representative video generation models reveals that diversity is highly dimension-specific: models with strong global diversity scores still collapse on specific factors, particularly Motion and Camera. These rankings persist after filtering unfaithful generations, indicating genuine capability differences rather than off-prompt outputs. Beyond measurement, controlled prompt interventions identify two fundamental bottlenecks: default mode convergence, where models fall back to dominant patterns under open-ended prompts; and realization gaps, where models fail to faithfully realize diverse, explicitly requested alternatives, particularly for temporal factors. The larger faithfulness losses for temporal factors highlight the difficulty of controlling motion and camera variation through text alone. DiVid thus shifts the study of diversity from measuring whether it exists to diagnosing where and why it collapses, and provides actionable directions for dimension-aware training objectives and control signals. The framework will be released to facilitate future research on diverse and controllable video generation.

1 Introduction

Recent advances in video generation have enabled models to produce visually compelling and increasingly realistic videos from text prompts (Ho et al., 2022; Brooks et al., 2024; Hu et al., 2026b). However, as these models become more aligned for high-quality generation, a limitation is becoming increasingly apparent: different samples generated from the same prompt often converge to highly similar outputs (illustrated in Figure 1). A model may repeatedly generate the same subject appearance, scene composition, camera trajectory, or motion pattern, despite the large space of possible valid interpretations. Such homogenization has been observed in large language models and text-to-image systems (Jiang et al., 2025; Han et al., 2025; Dombrowski et al., 2025; Teotia et al., 2025; Zhang et al., 2026). Video encodes spatiotemporal signals with far broader creative spaces, but its diversity remains poorly understood.

Diversity is a fundamental capability of generative models rather than merely an aesthetic preference. In practical applications, users often provide high-level intentions instead of fully specified scripts and expect generation systems to explore multiple plausible creative directions (Hu et al., 2022; Palani & Ramos, 2024; Davis et al., 2024; Choi et al., 2026). Limited diversity reduces the usefulness of generation systems by requiring excessive resampling, and may compress human creativity (Jiang et al., 2025) and amplify bias (Bianchi et al., 2023; Luccioni et al., 2023) at billion-user scale. More fundamentally, homogenization reveals where models fail to represent the underlying distribution of possible visual worlds (Zhu et al., 2024; Jena et al., 2025): a model that defaults to certain appearances, environments, or motion patterns may reflect biases in training data or objectives (Sun et al., 2026; Zhong et al., 2026; Chen et al., 2026). Long-tailed class-imbalanced training data causes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

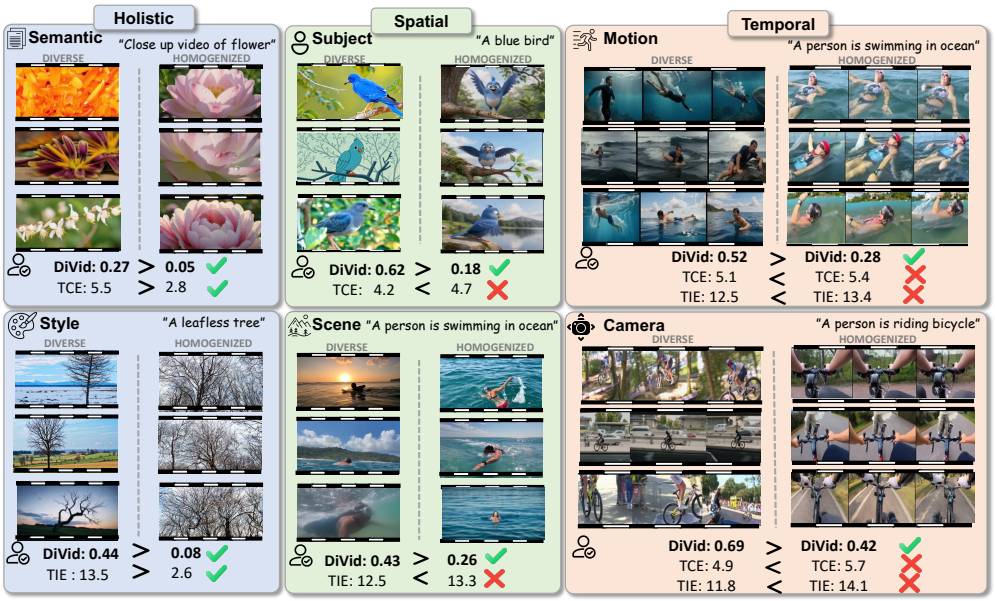

Figure 1: **Diagnosing dimension-specific homogenization**: For each dimension, the left set exhibits greater target-factor variation than the right set under the same prompt. DiVid consistently assigns the expected ordering in all dimensions, whereas global metrics such as TCE and TIE can fail to capture factor-level differences.

model collapse on tail concepts (Qin et al., 2023; Xu et al., 2024; Hong et al., 2026), while balanced data resampling degrades overall fidelity (Qin et al., 2023; Hu et al., 2026a). Targeted recovery thus requires locating the collapsed dimension.

Despite its importance, diagnosing diversity in video generation remains challenging. Existing video generation benchmarks primarily focus on quality, alignment and temporal consistency (Huang et al., 2024; Sun et al., 2025; Liu et al., 2024b). Beyond individual samples, recent diversity metrics (Zheng et al., 2025; Kazimi et al., 2026a) measure it through global scalars, which obscure fundamentally different failure modes. For example, a model may generate diverse scenes while producing identical camera motions, or vary subject appearance while maintaining the same visual style. These models may obtain similar overall diversity scores but require completely different solutions. Existing metrics thus answer *whether* diversity exists; advancing diversity requires diagnosing *where and how* it collapses.

In this work, we introduce **DiVid**, a dimension-resolved diagnostic framework for understanding diversity in video generation. Rather than a single global property, DiVid profiles diversity along six complementary and human-aligned dimensions: Semantic and Style diversity at the holistic level, and Subject, Scene, Motion, and Camera diversity at the factor level, each complemented by quality and faithfulness metrics to ensure that measured diversity is genuine. For each dimension, we design a dedicated extraction-and-comparison pipeline built from off-the-shelf computer vision tools, enabling reproducible, localized diagnosis of diversity strengths and failures.

We conduct extensive evaluations on representative video generation models. Our results reveal that diversity is highly dimension-specific: no model dominates across all dimensions, and different models exhibit distinct collapse patterns. Models with strong global diversity scores still collapse on temporal factors such as motion and camera. Lowering the guidance scale does not uniformly lift all dimensions, and diversity rankings disagree with quality and faithfulness rankings. These rankings persist after filtering unfaithful generations, confirming that they reflect genuine capability gaps rather than off-prompt variation. These findings demonstrate that existing global metrics can mask important capability differences between models.

Through controlled prompt interventions that manipulate individual dimensions, we identify two fundamental bottlenecks. First, *default-mode convergence*: when a target factor is unspecified or only broadly hinted, models converge to dominant patterns rather than exploring plausible alternatives.

Second, *realization gaps*: explicitly enumerating alternatives expands output span, but models do not consistently realize the requested candidates, particularly for Motion and Camera, echoing the guidance-scale findings above. Improving diversity thus requires dimension-aware objectives and control signals, not a single universal diversity metric.

Our contributions are summarized as follows: 1) We introduce **DiVid**, a dimension-level diagnostic framework that decomposes video diversity into six interpretable dimensions and provides localized analysis of diversity collapse; 2) We systematically evaluate representative video generation models and reveal that diversity is highly dimension-specific, with substantial gaps hidden by existing global metrics; 3) Through controlled intervention studies, we identify two fundamental diversity bottlenecks, default mode convergence and realization gaps, providing actionable directions for data curation, training objectives, and generation strategies.

2 Related Work

Video Generation Evaluation. Video generation benchmarks have progressed from general-purpose protocols (Liu et al., 2023; 2024b) to fine-grained dimensions (Huang et al., 2024; Sun et al., 2025; Zheng et al., 2025). However, these efforts predominantly score quality or faithfulness on a per-video basis. When diversity is considered, some works provide holistic global scores (Friedman & Dieng, 2023; Huang et al., 2024; Ibarrola & Grace, 2024; Zheng et al., 2025). Recent prompt-aware methods separate model-induced from prompt-induced variation (Ospanov et al., 2025; Jalali et al., 2026; 2025). However, they provide limited attribution to the subject, scene, or motion level, and thus lack the fine-grained diagnostic capability.

Diversity and Homogeneity in Generative Models. Homogenization in generative AI has been independently proved in text (Jiang et al., 2025), image (Han et al., 2025; Dombrowski et al., 2025), and video generation (Zhong et al., 2026; Sun et al., 2026), with alignment training further shown to worsen mode collapse (Liu et al., 2026a; Chen et al., 2026). Existing mitigation strategies expand the output distribution through prompt rewriting (Kazimi et al., 2026a;b), joint sampling (Liu et al., 2026b), or reward optimization (Liu et al., 2026a), yet predominantly rely on global objectives. While diversity evaluation has been studied in image generation (Han et al., 2025; Teotia et al., 2025; Astolfi et al., 2024; Zhang et al., 2026; Albuquerque et al., 2025), such systematic evaluation remains largely absent in video generation.

3 DiVid: Framework Design

3.1 Definition

Given a text prompt p, a video generation model m, and a set of videos $\mathcal{V}_{m,p} = \{v_i\}_{i=1}^{N}$ independently sampled under the same generation settings and different random seeds, we quantify how differently the samples realize prompt-compatible video solutions. Video diversity is a prompt-conditioned, set-level property, fundamentally distinct from single-video quality or instruction faithfulness. However, a video simultaneously encodes 4D spatiotemporal signals. A single holistic embedding can summarize overall dissimilarity but cannot localize where variation originates.

3.2 Six-Dimensional Diagnostic Pipeline

Guided by film theory and video representation studies (Bordwell & Thompson, 2013; Sun et al., 2023; Lin et al., 2025), a video can be perceived holistically through its content semantics and overall appearance, and be further decomposed along spatial composition (foreground subject, background scene) and temporal dynamics (subject motion, camera movement). Accordingly, we construct a two-level, six-dimensional diagnostic framework. The Holistic level (Semantic, Style) captures overall variation using established representation spaces; the Factor level further decomposes diversity into four dimensions: spatial composition (Subject, Scene) and temporal dynamics (Motion, Camera). Figure 2 illustrates the complete framework pipeline.

Design overview. All dimensions share a unified evaluation paradigm: extract a dimension-specific representation for each video, compute a pairwise similarity matrix, and aggregate it into a set-level

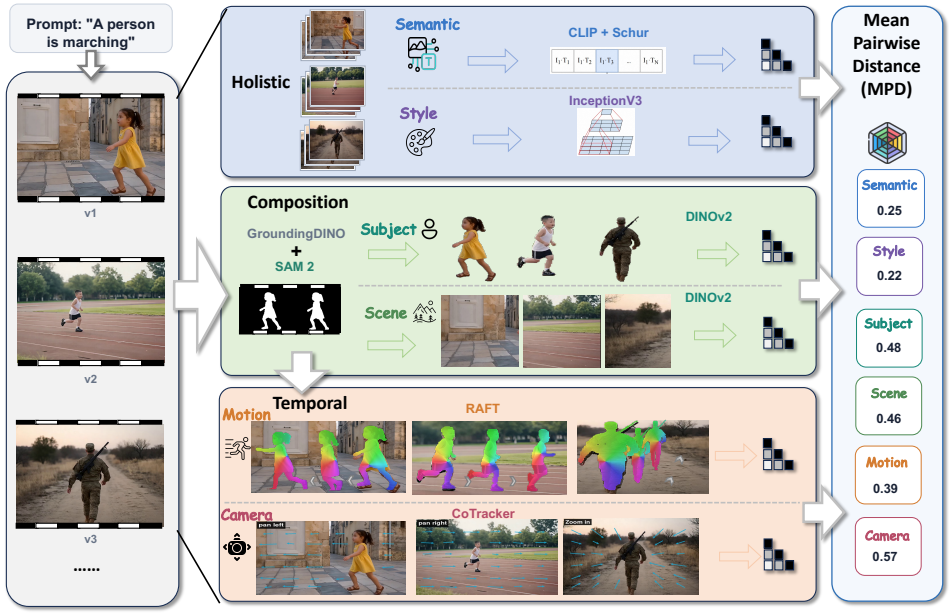

Figure 2: **Overview of the DiVid Framework**. Given a set of videos generated from a single prompt, our diagnostic framework extracts representations across six spatiotemporal dimensions to compute fine-grained diversity scores.

diversity score. The holistic Semantic and Style dimensions adapt image-level diversity kernels (Ospanov et al., 2025; Ibarrola & Grace, 2024) to video through frame-level aggregation; the remaining four factor-resolved dimensions introduce pipelines combining detection, segmentation, tracking, and motion decomposition to achieve dimension-level decoupled evaluation.

Holistic level. The **Semantic** dimension adopts the Schur-complement decomposition of the CLIP (Radford et al., 2021) kernel proposed by Scendi (Ospanov et al., 2025), subtracting the prompt-alignment component to capture semantic differences after removing prompt-shared information. The **Style** dimension aggregates frame-level InceptionV3 (Szegedy et al., 2016) representations to capture differences in overall appearance, color, lighting, texture, style, and layout.

Factor level: spatial composition. Subject and Scene rely on a subject localization pipeline: GroundingDINO (Liu et al., 2024a) detects the subject in a frame and SAM 2 (Ravi et al., 2025) propagates temporally consistent masks across frames. The **Subject** dimension extracts subject crops and encodes them with DINOv2 (Oquab et al., 2024) to measure differences in subject appearance. The **Scene** dimension masks the subject region and encodes the resulting background-focused frames with DINOv2 to capture environmental differences.

Factor level: temporal dynamics. This group disentangles two motion components: global camera movement and subject motion. The **Camera** dimension mainly fits a global affine transformation (translation, scaling, and rotation) to background trajectories extracted by CoTracker (Karaev et al., 2024), capturing differences in camera movement patterns and pace. The **Motion** dimension first compensates subject trajectories by removing estimated camera transform, then combines the resulting relative trajectories with canonicalized RAFT (Teed & Deng, 2020) optical flow to capture subject locomotion and internal deformation. This joint design minimizes leakage between camera movement and subject motion.

Metric Aggregation. Each dimension ultimately yields an $N \times N$ similarity matrix, which is aggregated using Mean Pairwise Distance (MPD) and Vendi Score (Friedman & Dieng, 2023). MPD serves as our primary metric for its interpretability and robustness to N, where larger values indicate more variation in the corresponding representation space. Vendi Score provides a complementary estimate of the effective number of distinct modes via eigenvalue entropy.

3.3 METRIC RELIABILITY

Human Evaluation. For diversity dimensions and faithfulness, we design separate blind human evaluation protocols. All six metrics correlate significantly with human judgments of set-level diversity (Spearman's $\rho = 0.531$–0.604, Kendall $\tau_b = 0.471$–0.539, all $p < 0.001$). Factor-level faithfulness metrics achieve 85.4%–90.3% directional agreement with human ratings. For all Diversity and Faithfulness metrics, the human evaluation agreement (Krippendorff's α) is greater than 0.667. These results support the perceptual validity of our diagnostic framework. Full protocols and statistics are in the Appendix B.4.

Framework Reliability. We validate the shared localization and each dimension through automatic validity filtering, human mask inspection, and strict-validity sensitivity analysis. The prompt–dimension invalid rates are only 0.83% for Subject and Scene, 0.76% for Motion, and 0.69% for Camera; invalid records are excluded rather than assigned zero. The corresponding four-dimension model-level rate ranges from 0% to 2.79%. Further human verification on 300 randomly sampled masks indicates that 93.3% contain neither salient foreground omission nor background leakage. Moreover, the dimension leaders remain unchanged when every model is required to retain at least 8 of 10 valid videos. Details and sensitivity checks are in Appendix B.3.

Diagnostic Utility. We prove the non-interchangeability and diagnostic value of the four factor-level dimensions through prompt-grouped cross-factor prediction, leave-one-factor-out diagnosis, and dimension-selective pair retrieval. The other three factors leave 58.1–73.4% of each held-out factor's variance unexplained; removing one factor changes the prompt-level top-model set for 19.5–26.7% of prompts. Each factor retrieves selective video pairs in 74.7–93.3% of the prompts. Thus, Subject, Scene, Motion, and Camera expose factor-specific, decision-relevant failures that a global metric would miss, detailed in Appendix B.5.

3.4 ADDITIONAL METRICS

Instruction Faithfulness (IF). It measures whether a video adheres to the prompt, ensuring valid diversity. A VLM (Qwen3.6-27B (Bai et al., 2025)) judge extracts the constraints for a single factor-level dimension (subject, scene, motion, or camera) from the prompt and independently assesses whether the generated video satisfies that dimension, assigning a score from 1 to 5, which is then linearly normalized to [0.2, 1]. The unconstrained dimension is marked N/A and Overall IF averages all applicable factor scores.

Quality. We employ Aesthetic Quality (AQ) and Imaging Quality (IQ) from VBench (Huang et al., 2024) to characterize visual fidelity, while the original VBench Temporal Flickering (TF) and Motion Smoothness (MS) metrics assess motion quality, ensuring that the observed diversity does not arise from artifacts or temporal degradation.

Reference diversity evaluation baselines. Following DPP-GRPO (Kazimi et al., 2026a), we report TCE and TIE as global set-level references that measure the spread of generated videos in CLIP and Inception feature spaces, respectively.

4 MAIN EVALUATION

4.1 EXPERIMENTAL SETUP

Prompt Construction. We construct our prompt set from two existing benchmarks to ensure generality, reproducibility, and broad coverage: VBench-Category (Huang et al., 2024) (8 general content categories), VBench-Dimension (10 quality dimensions), and T2V-CompBench (Sun et al., 2025) (7 categories of compositional constraints), sampling 10, 7, and 8 prompts per category respectively, for a total of 206 general prompts. All results reported are aggregated across the full prompt set; per-source breakdowns are provided in Appendix C.5 and show consistent trends.

Tested Models. Open-source models include Wan2.2-5B (Team Wan et al., 2025), Wan2.2-14B (Team Wan et al., 2025), CogVideoX1.5-5B (Yang et al., 2025), HunyuanVideo-1.5 (Wu et al., 2025).

Table 1: Faithfulness, quality and diversity profiles pooled over the 206 general prompts. The table is divided into open-source models, closed-source models, and two ablations. Results are macro-averaged over available prompt-level measurements. **Bold** and underlined denote the highest and second-highest values for normal setting.

	Faith / Quality					Prior Metrics		Holistic		Factor Level			
Model	IF	AQ	IQ	TF	MS	TCE	TIE	Semantic	Style	Subject	Scene	Motion	Camera
Wan2.2-5B (g=5)	0.848	0.543	0.672	**0.980**	0.989	**5.29**	**13.36**	**0.169**	**0.234**	**0.447**	**0.426**	0.326	0.097
Wan2.2-14B	0.915	**0.614**	0.699	0.972	0.983	4.54	13.29	0.144	0.224	0.412	0.394	0.344	0.148
CogVideo	0.839	0.516	0.622	0.973	0.985	4.12	12.66	0.142	0.214	0.414	0.377	0.323	0.045
Hunyuan	0.893	0.541	0.662	0.965	**0.991**	5.05	13.07	0.142	0.197	0.392	0.378	**0.433**	**0.490**
Wan2.7	0.963	0.563	0.722	0.972	0.985	4.94	12.74	0.158	0.201	0.397	0.390	0.413	0.303
HappyHorse	**0.986**	0.594	**0.729**	0.978	**0.991**	2.77	11.45	0.093	0.123	0.250	0.223	0.326	0.193
Seedance	0.955	0.595	0.667	0.970	0.990	5.16	13.09	0.160	0.162	0.380	0.364	0.425	0.463
Wan2.2-5B (g=3)	0.849	0.534	0.670	0.980	0.990	5.63	13.83	0.180	0.248	0.479	0.463	0.153	0.050
Wan2.2-5B (g=7)	0.875	0.549	0.680	0.979	0.989	4.79	13.22	0.147	0.229	0.428	0.402	0.141	0.048

Closed-source models include Wan2.7 (Alibaba Group, 2026), HappyHorse-1.5 (HappyHorse, 2026), and Seedance-1.5-pro (Team Seedance et al., 2025).

Evaluation Setup. For each prompt, each model independently generates 10 videos using distinct random seeds. All videos are generated at 5 seconds with a 16:9 aspect ratio, guidance scale 5.0 for open-source models. For Wan2.2-5B, guidance scales of 3.0 and 7.0 are evaluated for ablation. Other parameters are set to default, prompt extension disabled. All model-level statistics are computed with the prompt as the unit of analysis.

4.2 Dimension-Specific Diversity

As shown in Table 1, homogenization in video generation is dimension-specific and cannot be characterized by any single global score.

Diversity Ranks Reverse Across 6 Dimensions. Wan2.2-5B ranks first in Semantic, Style, Subject, and Scene, but falls into the bottom two in Motion and Camera. Hunyuan exhibits the opposite profile: it ranks first in both Motion and Camera, but only fourth to sixth across the four content dimensions. Similar reversals appear beyond the two leaders. Seedance ranks second in both temporal dimensions but sixth in Style, Subject, and Scene, whereas CogVideo ranks second in Subject but last in Motion and Camera. Overall, six of the seven models span at least three rank positions across the six dimensions. Prompt-paired bootstrap (20,000 resamples) and two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg correction confirm the central rank reversal, detailed in Appendix C.1.

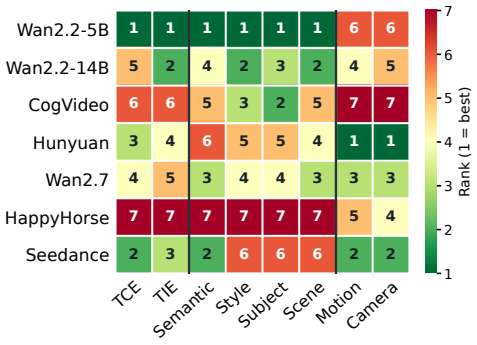

Figure 3: Rank Heatmap on all diversity metrics.

Global Metrics Conflate Opposite Diversity Collapse. Both TCE and TIE rank Wan2.2-5B first. These scores are close to Hunyuan's 5.05 and 13.07, yet their factor-level results are fundamentally different: Hunyuan achieves $1.33\times$ higher Motion MPD and $5.05\times$ higher Camera MPD, while Wan2.2-5B leads all four content dimensions. Additionally, Camera MPD spans a $10.8\times$ gap—from 0.490 (Hunyuan) to 0.045 (CogVideo). Thus, global metrics can summarize aggregate spread, but they neither preserve the full factor-level profile nor identify the source of homogenization. This diagnostic gap motivates factor-specific evaluation and optimization objectives.

Effect of CFG Scale. Lowering CFG from $g = 5$ to $g = 3$ increases Semantic, Style, Subject, and Scene diversity by 5.8–8.7%. Raising CFG to $g = 7$ improves IF, AQ, and IQ, respectively.

However, both deviations from $g = 5$ substantially reduce Motion and Camera diversity, indicating that CFG redistributes exploration across factors rather than uniformly controlling diversity. Motion and Camera require further control signals.

4.3 Joint Analysis on Quality, Faithfulness, and Diversity

No evaluated model performs strongly across faithfulness, quality, and all diversity dimensions. We analyze the three jointly: where they misalign, whether the measured diversity is genuine, and which model to choose under a stated preference.

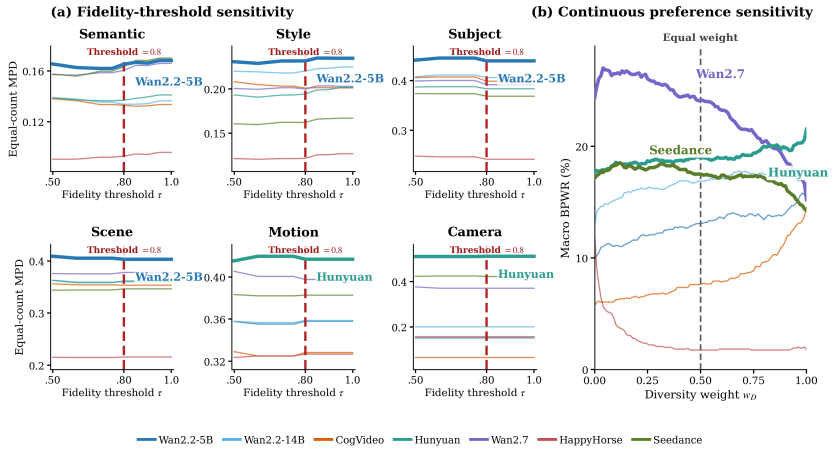

Figure 4: Faithfulness–Diversity trade-off. (a) Equal-count MPD ($k = 3$) under faithfulness threshold τ; colored curves highlight the Top-1 model at $\tau = 0.8$. (b) Six-dimension macro BPWR under diversity weight w_D.

Misalignment on Quality, Faithfulness and Diversity. Achieving high quality, instruction faithfulness, and rich diversity simultaneously remains a fundamental challenge in current video generation. HappyHorse ranks first in instruction faithfulness (IF=0.986) and imaging quality (IQ=0.729), yet ranks last in Semantic, Style, Subject, and Scene diversity. Conversely, Wan2.2-5B leads the four dimensions. At the prompt level, 35 of 49 significant correlations (after FDR correction) are negative (Table 14). Selecting a model from a quality leaderboard may favor the homogeneous one.

Validity of Measured Diversity. To test whether high MPD is a valid measurement or caused by off-prompt samples, we retain three faithful videos per model and prompt with faithfulness ≥ 0.8, which means they closely follow the prompt (Figure 4a). On the jointly valid prompt sets, dimension-level rankings remain unchanged: Wan2.2-5B still leads the four dimensions while retaining more than 98.3% diversity of its original MPD, and Hunyuan leads in Motion and Camera. The dimension-specific rankings therefore cannot be explained by unfaithful outputs or unequal pair counts. For quality, all models attain high TF and MS. Especially, Hunyuan ranks first in motion, camera and MS, so its diversity cannot be largely explained by artifacts or degraded motion smoothness.

Preference-Conditioned Model Selection. Sweeping the diversity weight w_D (Figure 4b) reveals no universally optimal model: Wan2.7 ranks first for $w_D \in [0, 0.83]$; Hunyuan takes over at $w_D \geq 0.84$. The overall recommendation shifts from Wan2.7 to Hunyuan as diversity receives greater weight. The best fidelity–diversity compromise depends on which dimensions matter most. The six-dimension macro curve averages win rates, not raw MPD, and serves as a rough recommendation summary. Details are in Appendix C.4.

5 Locating Diversity Bottlenecks via Prompt Intervention

Section 4 reports dimension-wise diversity profiles on general-purpose benchmark prompts. Here, we use controlled prompts that deliberately leave a target dimension open, isolating scenarios

Table 2: Mean MPD changes relative to the Specified condition; **bold** denotes the largest responses within each column, while Positive models reports the number of models with positive mean changes.

Model	Hinted				Omitted			Enumerated			
	Subject	Scene	Motion	Camera	Scene	Motion	Camera	Subject	Scene	Motion	Camera
Wan2.2-5B	+0.237	+0.014	**+0.128**	+0.032	+0.041	+0.049	+0.003	+0.450	+0.395	+0.168	+0.031
Wan2.2-14B	+0.203	+0.028	+0.100	+0.219	+0.034	-0.021	+0.009	+0.461	**+0.444**	+0.174	+0.167
CogVideo	+0.138	+0.056	+0.034	+0.010	+0.041	+0.010	+0.004	+0.411	+0.397	+0.095	+0.017
Hunyuan	+0.209	+0.075	-0.047	+0.081	+0.076	-0.020	+0.093	+0.492	+0.308	+0.067	+0.281
Wan2.7	+0.234	**+0.109**	+0.067	+0.398	**+0.105**	-0.004	+0.109	+0.468	+0.341	+0.146	+0.534
HappyHorse	+0.091	+0.030	+0.105	+0.295	-0.015	+0.015	+0.085	**+0.566**	+0.397	**+0.212**	+0.606
Seedance	**+0.269**	+0.080	+0.095	**+0.433**	+0.076	**+0.049**	**+0.340**	+0.460	+0.305	+0.159	**+0.614**
Mean	+0.197	+0.056	+0.069	+0.210	+0.051	+0.011	+0.092	+0.473	+0.370	+0.146	+0.321
Positive Models	7/7	7/7	6/7	7/7	6/7	4/7	7/7	7/7	7/7	7/7	7/7

where diversity in the target dimension is expected. This design distinguishes three behavioral regimes: default convergence under open prompts, diversity activation via generic cues, and candidate realization through explicit enumeration.

5.1 Experimental Setup

Prompt Construction. We use a Specified prompt that specifies all four dimensions as the paired baseline, then vary only one target dimension under three interventions with progressively increasing information (Table 3). **Specified** prompts ensure dimensional independence as much as possible: replacing one dimension does not constrain the others. **Omitted** removes the target-dimension description, exposing the model's unconstrained prior. **Hinted** replaces it with a dimension-level generic cue, testing whether a broad prompt can activate the dimension. **Enumerated** uses an LLM to enumerate prompts differing only in the target dimension, measuring model's controllability to realize predefined variation. All interventions share the same semantic core, distinguishing among settings with no cue, a broad cue, and explicit candidates. The experiment covers 20 semantic groups. Annotation details are in Appendix D.1.

Table 3: Illustration of the prompt settings.

Specified: *A golden retriever is resting on a beach, static camera.*

Dimension	Omitted	Hinted	Enumerated
Subject	N/A	*an animal*	{dog, cat, horse, ... }
Scene	remove *on a beach*	*somewhere*	{beach, forest, city, ... }
Motion	remove *resting*	*in motion*	{walks, runs, jumps, ... }
Camera	remove *static camera*	*camera moves*	{pan, zoom, dolly, ... }

Evaluation Metrics. For each model, semantic group, and target dimension, we compute the paired difference between the intervention MPD and the Specified MPD, denoted as Δ. Model-level results are the mean Δ over the 20 semantic groups. We also report directional consistency among models: 7/7 model agreement is robust, 6/7 is a strong trend.[1]

5.2 Default Mode Convergence

Open Prompts Trigger Default Preference. Table 2 shows that, although removing or explicitly relaxing a constraint is expected to increase diversity, every dimension still contains cases that revert to a default preference. For example, under Omitted, Motion improves for only four models, with a mean change of merely $+0.011$. Changing the generic wording gets consistent results: `in motion` and `performing a clear action` yield Motion improvements of $+0.069$ and $+0.054$, respectively, both with positive responses in $6/7$ models; similarly, `camera moves` and `with clear camera movement` yield Camera improvements of $+0.210$ and $+0.213$, respectively, both with $7/7$ model agreement. For Scene dimension, after

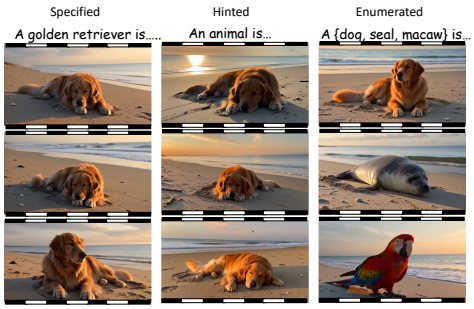

Figure 5: A subject convergence case: The model defaults "an animal" to "a golden retriever".

[1]MPD metrics across dimensions are in different representation spaces and are not directly comparable.

adding a generic cue, `somewhere` yields only a $+0.005$ improvement over Omitted. These results indicate that open constraints primarily activate high-probability default solutions rather than reflecting the natural variance of the open prompt.

Dimension-dependent Prompt Sensitivity. The response to prompt specificity differs by dimension: For Scene dimension, the generic cue `somewhere` improves over Omitted by only $+0.005$ in the mean row, whereas explicit enumeration lifts diversity by $+0.313$. The temporal dimensions instead respond gradually: `in motion` and `camera moves` improve over Omitted by $+0.058$ and $+0.118$, and Motion enumeration further gains $+0.077$. These profiles suggest that discrete, categorical dimensions such as Subject and Scene are unlocked only by explicit candidate enumeration, whereas temporal dimensions benefit from coupling high-level candidate planning with low-level execution constraints on trajectory and camera.

5.3 Realization Gap

Faithful Execution Bottleneck. Explicit enumeration consistently improves diversity across all four dimensions for all models. However, compared with the corresponding Hinted prompts, mean Faithfulness decreases in all dimensions, with larger decreases in Motion and Camera than in Subject and Scene (Figure 6a). Thus, larger MPD indicates candidate influence rather than faithful coverage; diversity and execution must be optimized jointly. Consequently, relying solely on text prompts to unlock fine-grained motion diversity is fundamentally bottlenecked, highlighting the necessity for future trajectory-based or multi-modal control signals to resolve temporal homogenization.

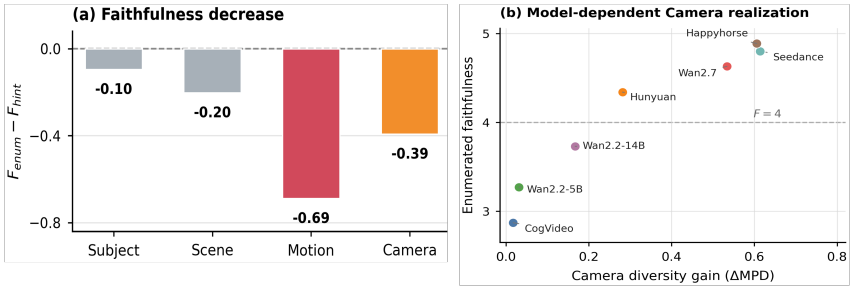

Figure 6: Faithful realization under enumeration on the original faithfulness scale. (a) Changes in target faithfulness from Hinted to Enumerated. (b) Camera diversity gain versus Enumerated faithfulness.

Model-dependent Responsiveness. Although enumeration obtains positive responses from all models, models differ substantially in realization faithfulness. In Figure 6b, Seedance, HappyHorse, and Wan2.7 all show large Camera gains, with high faithfulness, whereas CogVideo and Wan2.2-5B gain slightly with low faithfulness; the latter two also rank last on the main Camera Table 1. Furthermore, HappyHorse ranks last in Subject/Scene diversity while faithfully realizing the enumerated Subject and Scene candidates. The prompt-dependent default preference can likely be mitigated by prompt rewriting or decoding strategies, whereas capacity-limited failures call for better training data or architectural improvements. Our benchmark provides this model-level diagnosis.

6 Conclusion

We introduce **DiVid**, a reproducible dimension-level diagnostic framework that localizes diversity collapse in video generation across six interpretable dimensions. Systematic evaluations of representative models reveal distinct collapse patterns hidden by global metrics. Joint analysis on faithfulness and quality confirms that these differences reflect genuine generation capabilities rather than off-prompt variation. Beyond diagnosis, we find that CFG cannot uniformly control all dimensions, particularly temporal ones, open-ended prompts lead models toward default preferences, and explicitly requested alternatives remain difficult to realize faithfully, again most notably for Motion and Camera. DiVid helps identify promising future directions: dimension-aware training objectives and richer control signals beyond text prompts for achieving diverse, controllable and faithful video generation.

Ethics Statement

This research was conducted in strict adherence to the ICLR Code of Ethics. For human evaluation, the annotators we recruited possess a high level of education. They were fairly compensated for their time and effort in evaluation according to criteria. Our benchmark evaluates videos generated by publicly available T2V models using prompts from open-source benchmarks. Generated content may occasionally contain biased but unharmful content.

AI Use Statement

In this work, we used generative AI tools to generate the Enumerated intervention prompts, produce subject annotations, and as an automatic judge for the Instruction Faithfulness metric. We also used generative AI tools to polish the writing and assist in writing code. We have reviewed AI-assisted content: prompts were manually inspected before use; the automatic judge was validated by human evaluation; all AI-polished text and all AI-suggested citations are verified by the authors. We take full responsibility for the final content of this paper.

Reproducibility Statement

An anonymized implementation of the six-dimensional diversity evaluation framework is provided in the Supplementary Material; it computes diversity scores for a given set of generated videos using the same pipeline described in the paper. Additionally, we provide complete metric definitions and computation procedures in Appendix B.2, model versions and inference configurations in Section 4, and human evaluation details in Appendix B.4. All models used in this work are publicly available.

References

Isabela Albuquerque, Ira Ktena, Olivia Wiles, Ivana Kajić, Amal Rannen-Triki, Cristina Vasconcelos, and Aida Nematzadeh. Benchmarking diversity in image generation via attribute-conditional human evaluation. arXiv preprint arXiv:2511.10547, 2025. URL https://arxiv.org/abs/2511.10547.

Alibaba Group. Wan 2.7. Online model release, 2026. URL https://create.wan.video/wan2.7. Accessed September 25, 2026.

Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism Pareto fronts of conditional image generative models. arXiv preprint arXiv:2406.10429, 2024. URL https://arxiv.org/abs/2406.10429.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025. URL https://arxiv.org/abs/2511.21631.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1728–1738. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00175.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible Text-to-Image generation amplifies demographic stereotypes at large scale. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504. ACM, 2023. doi: 10.1145/3593013.3594095.

David Bordwell and Kristin Thompson. *Film Art: An Introduction*. McGraw-Hill, 10th edition, 2013.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, feb 2024. URL https://openai.com/index/video-generation-models-as-world-simulators/. Accessed September 25, 2026.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-Flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13–23, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Burgert_Go-with-the-Flow_Motion-Controllable_Video_Diffusion_Models_Using_Real-Time_Warped_Noise_CVPR_2025_paper.html.

Chubin Chen, Sujie Hu, Jiashu Zhu, Meiqi Wu, Jintao Chen, Yanxun Li, Nisha Huang, Chengyu Fang, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Taming Preference Mode Collapse via Directional Decoupling Alignment in Diffusion Reinforcement Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12775–12786. IEEE, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Taming_Preference_Mode_Collapse_via_Directional_Decoupling_Alignment_in_Diffusion_CVPR_2026_paper.html.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.137.

DaEun Choi, Kihoon Son, Jaesang Yu, HyunJoon Jung, and Juho Kim. IdeaBlocks: Expressing and reusing divergent intents for graphic design exploration using generative AI. In *Proceedings of the ACM Designing Interactive Systems Conference*, pp. 621–642. ACM, 2026. doi: 10.1145/3800645.3813005.

Richard Lee Davis, Thiemo Wambsganss, Wei Jiang, Kevin Gonyop Kim, Tanja Käser, and Pierre Dillenbourg. Fashioning creative expertise with generative AI: Graphical interfaces for design space exploration better support ideation than text prompts. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 1–26. ACM, 2024. doi: 10.1145/3613904.3642908.

Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3029–3039. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00288.

Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2VControl: Disentangled and unified video motion synthesis control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14051–14060, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Feng_I2VControl_Disentangled_and_Unified_Video_Motion_Synthesis_Control_ICCV_2025_paper.html.

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=g97OHbQyk1.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–12, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Geng_Motion_Prompting_Controlling_Video_Generation_with_Motion_Trajectories_CVPR_2025_paper.html.

Evans Xu Han, Alice Qian Zhang, Haiyi Zhu, Hong Shen, Paul Pu Liang, and Jane Hsieh. POET: Supporting prompting creativity and personalization with automated expansion of Text-to-Image generation. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pp. 1–18. ACM, 2025. doi: 10.1145/3746059.3747710.

HappyHorse. HappyHorse 1.5. Online model release, 2026. URL https://happyhorse15.com/. Accessed September 25, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646. Curran Associates, Inc., 2022. doi: 10.52202/068431-0628.

Feng Hong, Jiangchao Yao, Yifei Shen, Dongsheng Li, Ya Zhang, and Yanfeng Wang. Improving diffusion models for class-imbalanced training data via capacity manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=wSGle6ag5I.

Junyan Hu, Wei Luo, Tong Chen, Xiaobao Yang, and Zhiqiang Hou. PD-CBDM: Training class-balancing diffusion models with perceptual distinguish loss. *Mathematics*, 14(10):1576, 2026a. doi: 10.3390/math14101576.

Teng Hu, Jiangning Zhang, Hongrui Huang, Ran Yi, Zihan Su, Jieyu Weng, Zhucun Xue, Lizhuang Ma, Ming-Hsuan Yang, and Dacheng Tao. Evolution of video generative foundations. arXiv preprint arXiv:2604.06339, 2026b. URL https://arxiv.org/abs/2604.06339.

Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make It Move: Controllable Image-to-Video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18198–18207. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01768.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060.

Francisco Ibarrola and Kazjon Grace. Measuring diversity in co-creative image generation. In *Proceedings of the 15th International Conference on Computational Creativity*, pp. 254–261. Association for Computational Creativity, 2024.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. In *Advances in Neural Information Processing Systems*, volume 38, pp. 132953–132990. Curran Associates, Inc., 2025. doi: 10.52202/085713-4005.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional Vendi score: Prompt-aware diversity evaluation for generative AI models and LLMs. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1369–1377. PMLR, 2026. URL https://proceedings.mlr.press/v300/jalali26a.html.

Rohit Jena, Ali Taghibakhshi, Sahil Jain, Gerald Shen, Nima Tajbakhsh, and Arash Vahdat. Elucidating optimal Reward-Diversity tradeoffs in Text-to-Image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 232–242. IEEE, 2025. doi: 10.1109/WACV61041.2025.00033.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial Hivemind: The Open-Ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, volume 38, pp. 90705–90774. Curran Associates, Inc., 2025. doi: 10.52202/085713-2732.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *Proceedings of the European Conference on Computer Vision*, pp. 18–35. Springer, 2024. doi: 10.1007/978-3-031-73033-7_2.

Tahira Kazimi, Connor Dunlop, and Pinar Yanardag. Diverse Video Generation with Determinantal Point Process-Guided Policy Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12839–12848. IEEE, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kazimi_Diverse_Video_

Generation_with_Determinantal_Point_Process-Guided_Policy_Optimization_CVPR_2026_paper.html.

Tahira Kazimi, Heather Yu, and Zhiqiang Lao. DIVA: Diverse video generation with agents. In *Proceedings of the IEEE Conference on Artificial Intelligence*, pp. 1226–1231. IEEE, 2026b. doi: 10.1109/CAI68641.2026.11536189.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, fast, and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00981.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Yu Tong Tiffany Ling, Yuhan Huang, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. Towards understanding camera motions in any video. In *Advances in Neural Information Processing Systems*, volume 38, pp. 148814–148857. Curran Associates, Inc., 2025. doi: 10.52202/085713-4477.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1864–1873. IEEE, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Liu_DiverseGRPO_Mitigating_Mode_Collapse_in_Image_Generation_via_Diversity-Aware_GRPO_CVPR_2026_paper.html.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pp. 38–55. Springer, 2024a. doi: 10.1007/978-3-031-72970-6_3.

Xinshuang Liu, Runfa Blark Li, and Truong Nguyen. Consistency-preserving diverse video generation. arXiv preprint arXiv:2602.15287, 2026b. URL https://arxiv.org/abs/2602.15287.

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*, pp. 22139–22149. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.02090.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for Fine-Grained evaluation of Open-Domain Text-to-Video generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62352–62387. Curran Associates, Inc., 2023. doi: 10.52202/075280-2723.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 56338–56351. Curran Associates, Inc., 2023. doi: 10.52202/075280-2458.

Zijie Meng. OrthoMotion: Disentangling camera and subject motion via geometry-semantics orthogonal attention. arXiv preprint arXiv:2606.22835, 2026. URL https://arxiv.org/abs/2606.22835.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=j7kdXSrISM.

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. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Azim Ospanov, Mohammad Jalali, and Farzan Farnia. Scendi score: Prompt-aware diversity evaluation via Schur complement of CLIP embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16927–16937. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01572.

Srishti Palani and Gonzalo Ramos. Evolving roles and workflows of creative practitioners in the age of generative AI. In *Proceedings of the ACM Conference on Creativity and Cognition*, pp. 170–184. ACM, 2024. doi: 10.1145/3635636.3656190.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-Balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18434–18443. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01768.

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 *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M Weber. Guidance in the frequency domain enables high-fidelity sampling at low CFG scales. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0cdXElXkk6.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416. IEEE, 2025. doi: 10.1109/CVPR52734.2025.00787.

Mingzhen Sun, Weining Wang, Xinxin Zhu, and Jing Liu. MOSO: Decomposing MOtion, scene and object for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18727–18737. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01796.

Yongxu Sun, Michael Saxon, Ian Yang, Anna-Maria Gueorguieva, and Aylin Caliskan. VEAT quantifies implicit associations in text-to-video generator Sora and reveals challenges in bias mitigation. arXiv preprint arXiv:2601.00996, 2026. URL https://arxiv.org/abs/2601.00996.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826. IEEE, 2016. doi: 10.1109/CVPR.2016.308.

Team Seedance, Heyi Chen, Siyan Chen, Xin Chen, Yanfei Chen, Ying Chen, Zhuo Chen, Feng Cheng, Tianheng Cheng, Xinqi Cheng, et al. Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507, 2025. URL https://arxiv.org/abs/2512.13507.

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. URL https://arxiv.org/abs/2503.20314.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24.

Revant Teotia, Candace Ross, Karen Ullrich, Sumit Chopra, Adriana Romero-Soriano, Melissa Hall, and Matthew Muckley. DIMCIM: A quantitative evaluation framework for default-mode diversity and generalization in text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16431–16440. IEEE, 2025. doi: 10.1109/ICCV51701.2025.01525.

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. ATI: Any trajectory instruction for controllable video generation. arXiv preprint arXiv:2505.22944, 2025. URL https://arxiv.org/abs/2505.22944.

Haoyuan Wang, Yabo Chen, Haibin Huang, Chi Zhang, and Xuelong Li. Directing the world: Fast autoregressive video generation with compositional human-camera control. arXiv preprint arXiv:2606.27964, 2026. URL https://arxiv.org/abs/2606.27964.

Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013. doi: 10.1109/ICCV.2013.441.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11. Association for Computing Machinery, 2024. doi: 10.1145/3641519.3657518. URL https://doi.org/10.1145/3641519.3657518.

Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. arXiv preprint arXiv:2511.18870, 2025. URL https://arxiv.org/abs/2511.18870.

Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. Rethinking noise sampling in class-imbalanced diffusion models. *IEEE Transactions on Image Processing*, 33:6298–6308, 2024. doi: 10.1109/TIP.2024.3485484.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Xiaofeng Zhang, Aaron Courville, Michal Drozdzal, and Adriana Romero-Soriano. The intricate dance of prompt complexity, quality, diversity and consistency in T2I models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=RBIBMCdw7y.

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. URL https://arxiv.org/abs/2503.21755.

Haonan Zhong, Wei Song, Tingxu Han, Maurice Pagnucco, Jingling Xue, and Yang Song. FAIRT2V: Training-free debiasing for text-to-video diffusion models. arXiv preprint arXiv:2601.20791, 2026. URL https://arxiv.org/abs/2601.20791.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. Is Sora a world simulator? a comprehensive survey on general world models and beyond. arXiv preprint arXiv:2405.03520, 2024. URL https://arxiv.org/abs/2405.03520.

A Discussion and Future Directions

A.1 Dimension Extraction

The four factor dimensions (Subject, Scene, Motion, Camera) depend on automatic extraction of video constituents. Accurate extraction is an independent line of work: subject detection and segmentation (Liu et al., 2024a; Ravi et al., 2025), point tracking and optical flow (Karaev et al., 2024; Teed & Deng, 2020), and camera-motion understanding and pose estimation for video (Lin et al., 2025; Li et al., 2025). Recovering motion and camera information from 2D frames remains the hardest part (Lin et al., 2025; He et al., 2025). Our aim is not more accurate extraction, but natural disentanglement of video constituents and set-level measurement of diversity. Every score is a pairwise similarity computed within the videos of the same prompt, aggregated at the prompt level, and used for relative ranking across models; Relative discriminability, reproducibility, and efficiency are what matter here. Appendix B.3 establishes the validity and robustness of the metrics through explicit validity controls, comprehensive data-quality statistics, human verification, and sensitivity analyses.

Our extraction accordingly follows the background-compensation paradigm validated in action recognition (Wang & Schmid, 2013), adapted to set-level measurement: we exclude the foreground with a dilated mask and fit a global partial-affine transform (translation, rotation, uniform scale) by RANSAC, which is richer than subtracting an averaged background vector from an averaged foreground vector (Sun et al., 2025); Motion adds a canonicalized deformation channel with two interpretable kernels for relative displacement and internal deformation; Camera aggregates translation, scale, and rotation into a 16-D descriptor; All key parameters are carefully selected by pilot analyses to reflect its intended factor and are then fixed across all models and prompts. The pipeline is fully open-source and automatic, agrees closely with human judgment (Table 5), and makes dimension-resolved diversity measurement efficient at benchmark scale. As motion and camera estimation for generated video improves in reliability and efficiency, we can update the corresponding modules.

A.2 Toward Dimension-Aware Diversity Enhancement

DiVid points to three complementary directions for diversity enhancement: data and training objectives determine which variations a model learns, control interfaces determine how users specify the desired variation, and inference strategies determine how to realize it.

- **Data and Objectives.** Section 5 shows that models often converge to dominant defaults. For example, HappyHorse repeatedly interprets *an animal* as a golden retriever, whereas explicitly listing candidates expands its Subject diversity. Meanwhile, the real-video retrieval reference in Appendix C.2 leads in the four content-related dimensions but ranks only fourth in Motion and Camera, suggesting that broad web-video coverage alone does not guarantee equally rich temporal variation. Prior image-generation studies similarly show that class imbalance causes tail-mode collapse, while global rebalancing can reduce overall fidelity (Qin et al., 2023; Hu et al., 2026a). Future methods should therefore curate data based on the diagnosed collapse dimensions and develop factor-specific training objectives that improve coverage and instruction following, rather than uniformly rebalancing the training distribution.
- **Control Interfaces.** Prompt interventions show that different dimensions require different forms of control. Explicitly listing candidates effectively expands discrete content factors, but produces less reliable gains in Motion and Camera. Text alone is therefore insufficient for precise control. Future systems could adopt hierarchical control, using text-level candidate specification for content factors and structured conditions for Motion and Camera, such as trajectories, keypoints, and camera paths.
- **Inference and Sampling.** In Section 4.2, we systematically examine how CFG scale affects each diversity dimension. As a preliminary feasibility check, we try adapting three diversity-enhancement methods from image generation to video: a bell-shaped CFG schedule; frequency-decoupled guidance with $g{=}1.5$ for low frequencies and $g{=}5.0$ for high frequencies (Sadat et al., 2026), and condition-annealed sampling that adds Gaussian noise to the T5 text embedding early in sampling and gradually reduces it to zero (Sadat et al., 2024). In pilot experiments, these methods do not produce consistent improvements across dimensions: some apply only to specific models, while others may trade quality for diversity. This exposes a common limitation of scalar guidance: it cannot select which factor to diversify. Future inference methods should instead use

dimension-specific guidance and factor-aware sampling to explore the target dimension while jointly preserving quality and faithfulness.

These three directions share one conclusion: a global diversity metric cannot provide a concrete optimization target, making dimension-level diagnosis essential for targeted enhancement. Recent controllable video-generation methods use structured camera poses, trajectories, or flow to control camera and object motion (He et al., 2025; Wang et al., 2024; Geng et al., 2025; Burgert et al., 2025; Wang et al., 2025; Feng et al., 2025), with newer approaches further disentangling these factors or composing their control over longer horizons (Meng, 2026; Wang et al., 2026). These advances provide practical mechanisms for acting on our diagnosis. DiVid complements them by locating the collapsed factor, evaluating whether a control method expands its intended dimension, and providing a low-cost automatic signal for training or sampling while jointly monitoring quality and faithfulness.

B Additional Details of the Framework

This appendix provides robustness of generation number, implementation details, validity checks, and scope conditions that complement Section 3.

B.1 Robustness of Generation Number

To verify that generating 10 videos per prompt is sufficient, we randomly select 10 prompts and generate 30 videos for each. In Figure 7, the models retain the same six-dimensional ordering from $N = 10$ to $N = 30$ on MPD and Vendi. The MPD scores also remain stable. Since Vendi Score is equivalent to the effective number of distinct samples, the raw score increases with set size but at a diminishing rate, while model ordering remains stable. Figure 8 visualizes the intra-model and inter-model homogenization by PCA.

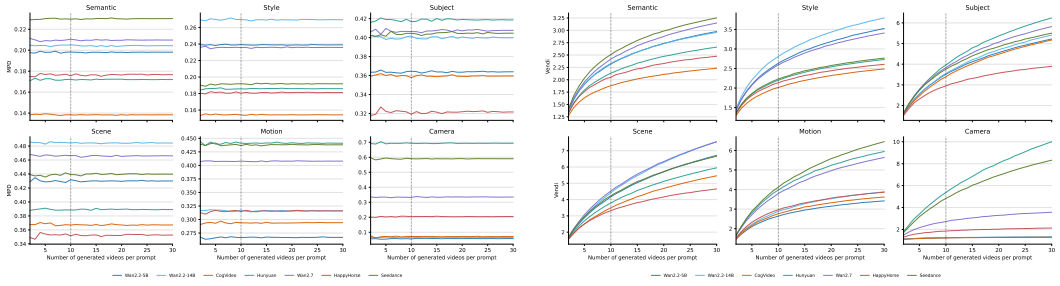

Figure 7: Sensitivity to the number of generated videos. Left: MPD. Right: raw Vendi Score.

B.2 Additional Information about DiVid Framework

Each group of 10 videos takes approximately 30 seconds for the full six-dimension evaluation on one A100 GPU. All dimensions follow the same set-level interface: each video is mapped to a dimension-specific representation, the representations form a pairwise kernel, and the kernel is aggregated over the videos generated for the same prompt. Encoders and geometric hyperparameters are fixed for all models.

B.2.1 Details on Diversity Metrics

TCE and TIE references. For comparison with prior global diversity metrics, we follow the video protocol of TCE/TIE (Kazimi et al., 2026a). Since truncated covariance entropy is sensitive to the number of samples and retained eigenvalues, our TCE/TIE values, computed from ten videos per prompt, are used only for within-benchmark comparison and are not numerically comparable to prior results based on twenty-video sets.

Semantic. We uniformly sample eight frames from each video and encode them with CLIP ViT-L/14. After frame averaging and ℓ_2 normalization, the video embeddings form X, while the normal-

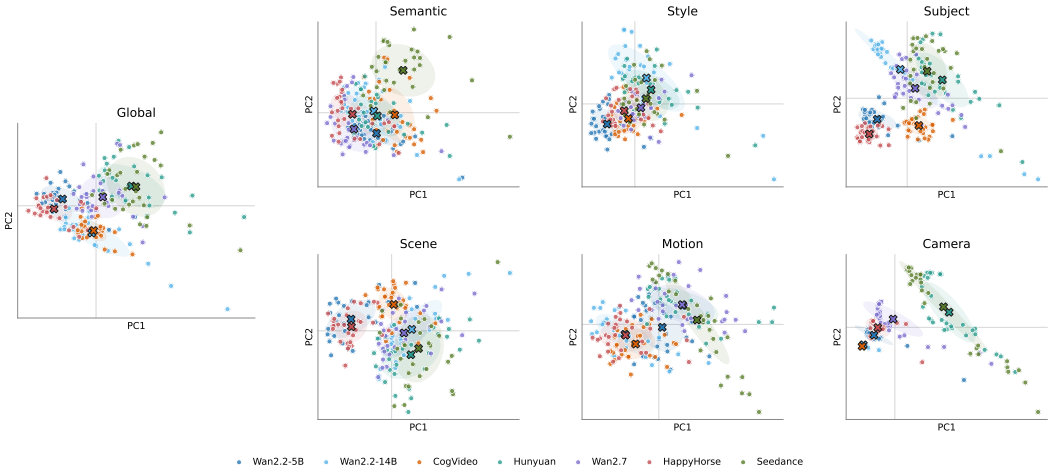

Figure 8: Each point represents a generated video. PCA visualization illustrates both intra-model and inter-model homogenization across six diversity dimensions and the global profile.

ized prompt embedding is z. We remove the prompt-alignment direction using

$$K_{\text{sem}} = XX^\top - (Xz)(Xz)^\top. \tag{1}$$

Negative eigenvalues caused by numerical error are clipped, and the kernel is normalized by its diagonal. The pairwise distance is $1 - K_{ij}$. This operation reduces shared prompt alignment but does not remove all semantic dependence on the prompt.

Style diversity. We uniformly sample eight frames per video and extract 2048-dimensional InceptionV3 pool3 features. Frame features are individually ℓ_2-normalized, averaged over time, and normalized again to form a video-level descriptor. Pairwise cosine similarities define the style kernel, whose mean pairwise distance measures diversity in global appearance, including composition, texture, color, and lighting.

Subject diversity. GroundingDINO searches up to 32 frames for an initialization subject box, and SAM2 propagates the mask forward and backward. A track requires at least three valid mask frames. Multi-subject videos are processed per subject and aggregated using mean mask area as the weight. For each valid frame, we crop the tight subject box and replace pixels outside the mask with the frame-wise RGB mean to reduce background leakage. DINOv2 [CLS; mean patch] features are averaged over time and normalized before the pairwise kernel is computed.

Scene diversity. We suppress the annotated foreground using the union of the tracked subject masks, replacing subject pixels with a Gaussian-blurred version of the frame (radius 12). Each background-focused frame is represented by the concatenation of the DINOv2 class token and the mean patch token. Frame features are averaged over time and ℓ_2-normalized to obtain a video descriptor. Pairwise cosine similarities define the scene kernel, and scene diversity is measured by its mean pairwise distance.

Motion diversity. CoTracker is initialized at four temporal anchors with a 10×10 subject grid and a 12×12 background grid. For each adjacent frame pair, a robust global transform H_t is fitted from background tracks, and subject motion is compensated as

$$r_{p,t} = x_{p,t+1} - H_t(x_{p,t}). \tag{2}$$

A relative similarity transform fitted to the compensated subject tracks yields translation, scale, and rotation signals. Their temporal statistics, path characteristics, and an eight-bin direction histogram form a 24-dimensional trajectory descriptor. In parallel, smoothed subject boxes are mapped to a 256×256 canonical canvas. RAFT flow is computed between canonicalized frames, and the fitted rigid similarity flow is removed. The temporal means and standard deviations of residual magnitude,

moving ratio, and an eight-bin direction histogram form a 20-dimensional deformation descriptor. After dimension-wise normalization by fixed scales, the two descriptors define RBF kernels with $\sigma = 3.0$, which are combined as

$$K_{\text{motion}} = 0.5K_{\text{traj}} + 0.5K_{\text{def}}. \tag{3}$$

Only videos with valid trajectory and deformation estimates are included.

Camera diversity. Background points outside a five-pixel dilation of the subject mask are used to fit a partial-affine transform. A homography replaces it when sufficiently supported and substantially more accurate, with its center-local motion converted to the same camera parameters. Each frame pair yields normalized horizontal translation, vertical translation, log scale, and rotation. Mean, standard deviation, mean absolute magnitude, and the 90th percentile over time produce a 16-dimensional descriptor; frozen feature scales and an RBF kernel with $\sigma = 2.0$ are used for all models.

B.2.2 Subject Annotation

We define the subject as the foreground entity in a video that is clearly visible, recognizable, and suitable for stable localization and segmentation. By default, we retain only the single most central subject. Multiple subjects are annotated only when the prompt contains clearly coordinated core entities. Subject labels are normalized to generic, single-word category nouns (e.g., `person`, `dog`, and `car`), without modifiers such as color, material, age, or style, and without proper names or character names. Nouns appearing in prepositional phrases, supporting objects, and background reference objects are usually not annotated as subjects. We use GPT-5 to produce the initial subject annotations for all 206 prompts, and the annotation instruction is shown in Figure 19. For the prompts in Section 5, the subject labels are predefined during prompt construction. After the initial annotation, we asked three annotators to manually review all subject labels. For 94.7% of the prompts, the GPT-5 annotation was unanimously accepted by all three annotators. For the remaining prompts with disagreement, we adopted a majority-vote strategy and used the label preferred by most annotators as the final subject annotation.

B.2.3 Set-Level Aggregation

For a prompt–model set with N valid videos, Mean Pairwise Distance is

$$\text{MPD}(K) = \frac{2}{N(N-1)} \sum_{i<j} (1 - K_{ij}). \tag{4}$$

It is the expected distance between two uniformly sampled candidates from the observed set. Because each dimension uses a different representation and kernel, raw MPD values are compared only within the same dimension.

Vendi Score normalizes the kernel eigenvalues to sum to one and computes $\exp(-\sum_j \lambda_j \log \lambda_j)$. It provides an effective-mode perspective, whereas MPD is the primary metric because its pairwise interpretation is direct.

B.3 Validity Checks and Scope

Automatic validity controls. GroundingDINO scans up to 32 sampled frames for an initialization box, using box and text thresholds of 0.30 and 0.25. SAM 2 then propagates each detected subject mask both forward and backward. A video is retained only if every annotated subject has at least three mask frames whose area lies between 0.1% and 95% of the image. Multi-subject instances are tracked separately before aggregation. At the prompt level, a factor is excluded rather than assigned zero when fewer than two videos, or fewer than 30% of the generated videos, retain valid subject masks. Subject and Scene use complementary leakage controls. For Subject, pixels outside the mask are replaced by the frame-wise RGB mean before the tight crop is encoded. For Scene, the masked foreground is replaced by a Gaussian-blurred region with radius 12. For temporal metrics, Camera samples background points outside a five-pixel dilation of the subject mask, while Motion samples subject points from an eroded interior region. This exclusion margin reduces boundary contamination in both directions.

Table 4: Motion sensitivity pilot. Each row aggregates three prompts that share a subject and differ in the indicated motion property.

Group	Controlled variation	Motion MPD
1	Breathing / ear twitch / deep breath	0.254
2	Walk left / right / toward camera	0.469
3	Slow / normal / fast locomotion	0.483
4	Sleep / stand and stretch / chase	0.616
5	Shake / scratch / dig	0.456

Robust temporal estimation. CoTracker is initialized at four temporal anchors using a 10×10 subject grid and a 12×12 background grid. Geometry estimation requires at least 12 subject points and 20 background points. Background transformations are fitted with RANSAC and rejected when confidence is below 0.35; subject trajectories additionally require an inlier ratio of at least 0.50. A Camera or trajectory descriptor is retained only when at least two frame pairs and at least 25% of the attempted pairs are valid. Motion is accepted only when both the background-compensated trajectory and canonical-deformation branches pass their validity checks.

Statistics. Among all videos, all annotated subjects receive an initialization detection in 98.85% of cases; conditional on detection, SAM 2 execution succeeds in 99.94%, and 96.60% of all expected videos pass the complete mask-validity gate. Among attempted temporal estimates, 96.70% of background-transform pairs, 94.21% of background-valid subject-trajectory pairs, and 98.48% of canonical-deformation pairs pass their respective gates. At the final prompt–dimension level, invalid rates are 0.83% for Subject and Scene, 0.76% for Motion, and 0.69% for Camera; the model-level rate across these four dimensions ranges from 0% to 2.79%. Invalid records are treated as missing and paired comparisons use only the intersection of valid prompts. Requiring every model to retain at least 8 of 10 valid videos still preserves the leading model in each factor dimension.

Human mask verification. To evaluate errors that cannot be detected by internal status flags, we randomly sampled 300 automatically accepted mask overlays across models and prompts. The text prompt and mask overlay were shown to three annotators, with 100 samples assigned to each annotator. A mask passed only when the target subject had no salient omission and the background or other objects had no salient leakage. The aggregated pass rate was 93.3%; the remaining failures were concentrated in heavy occlusion and multi-object scenes.

Motion sensitivity. A controlled 15-prompt set varies only the motion description of the same animal subject. For example, subtle-motion prompts include ”a bear sleeping with its body rising and falling gently” and ”a bear sleeping while its ears twitch occasionally”; direction variants include ”a bear walking to the left,” ”walking to the right,” and ”walking toward the camera”; larger action changes range from ”a bear sleeping peacefully” to ”standing up and stretching” and ”running while chasing a salmon.” In Table 4, motion MPD is lowest for subtle breathing and local twitches, increases for direction or speed changes, and is highest when the action type and amplitude change together.

Metric scope. Subject, Scene, Motion, and Camera require an explicit, localizable subject. Most prompts in our benchmark specify a focal entity, covering the vast majority of the evaluated prompts and common text-to-video use cases. Prompts outside this scope are excluded by the explicit validity checks.Motion and Camera are 2D image-plane measurements. The partial-affine Camera descriptor captures dominant global shot motion rather than full 3D camera pose. Strong parallax, dynamic backgrounds, or shot cuts may require richer geometric models, while Motion remains sensitive to tracking quality and does not assess physical correctness. We therefore use these dimensions to diagnose visible motion and camera diversity at the model level.

Representation. As with other feature-based diversity metrics, DiVid measures variation through specific encoders. We use representations matched to each dimension and validate them meticulously. MPD values are therefore interpreted and compared within each dimension rather than across dimensions. The framework is modular: alternative encoders can be substituted while retaining the same factor definitions, pairwise aggregation, and evaluation protocol.

B.4 HUMAN EVALUATION

Evaluation Protocol. We recruited six trained annotators and randomly sampled 50 prompts from different categories. For each prompt, three disjoint model pairs were randomly drawn from all seven models, producing 150 unique prompt–model-pair items. Each pair was judged on six diversity dimensions using a five-level ordered scale: *A clearly more diverse* / *A slightly more diverse* / *about the same* / *B slightly more diverse* / *B clearly more diverse*. Each annotator evaluated 40 model pairs, resulting in 240 valid judgments that ensure full coverage of all 150 items, with 90 items evaluated by two annotators to allow for inter-annotator agreement analysis. Model identity, file names, left/right position, and within-set video order were hidden or randomized throughout. Annotators also rated on factor-level faithfulness across four dimensions (Subject, Scene, Motion, and Camera). Specifically, given the prompt's description for each factor, annotators judged whether the generated videos faithfully followed the corresponding constraint on a 1–5 scale; N/A was selected when the prompt did not specify that factor. Strict requires the automatic and human scores to be identical, while Direction counts agreement when both scores are positive (4–5), both are negative (1–2), or both are N/A.

Quality Control. All annotators hold at least a bachelor's degree and completed a unified training session covering the judgment criteria for each dimension before the formal evaluation began. We explicitly train annotators to focus solely on diversity variations in the specific dimension, avoiding confusion. For motion-related dimensions (Motion and Camera), the annotation interface enforced a minimum playback requirement: annotators must have watched at least 80% of the total video duration before submission was enabled. Appearance dimensions (Subject, Scene) and motion dimensions (Motion, Camera) were presented on separate screens to minimize cross-dimension interference. To monitor annotation reliability, we inserted 2–3 gold standard items per batch of 40 judgments. Each gold item consisted of one set containing 2–3 identical duplicate videos, and the other contained 10 manually curated, clearly distinct videos spanning diverse content. Across all annotators and gold items, every response correctly identified the set of distinct videos as clearly more diverse. These checks confirm that annotators remained attentive and applied the diversity criteria consistently throughout the study.

Table 5: Human evaluation for 6-dimension diversity. All correlations have $p < 0.001$.

Dimension	Spearman ρ	Kendall τ_b	Kripp. α
Semantic	0.604	0.539	0.703
Style	0.563	0.522	0.682
Subject	0.549	0.503	0.679
Scene	0.541	0.489	0.680
Motion	0.531	0.471	0.692
Camera	0.556	0.502	0.686

Table 6: Human evaluation for factor-specific faithfulness.

Factor	Strict	Direction	Kripp. α
Subject	63.4%	85.4%	0.699
Scene	65.4%	88.9%	0.705
Motion	63.7%	90.1%	0.742
Camera	69.3%	90.3%	0.766

Results. In Table 5 and 6, all six diversity metrics correlate positively with human pairwise preferences, with Spearman $\rho = 0.531$–0.604 and Kendall $\tau_b = 0.471$–0.539 ($p < 0.001$). Factor-faithfulness direction agreement is 85.4%–90.3%. Krippendorff's α exceeds the commonly accepted reliability threshold of 0.667. The results support model-level ranking and diagnostic use.

B.5 Utility of Factor Level Metrics

We test whether each factor retains information that the other three cannot recover without claiming that the factors are statistically independent or form a unique taxonomy. The analysis uses the 195 prompts valid for all four factors and all seven models (1,365 model–prompt records). Within each prompt, MPD values are converted to seven-model rank percentiles. Held-out prediction uses prompt-grouped 10-fold cross-validation, so records from one prompt never appear in both train and test folds.

Table 7: Incremental diagnostic utility. Grouped-CV predicts one factor from the other three. Top-model change compares the prompt-level top set before and after deleting the factor.

Factor	VIF	Grouped-CV R^2 (95% CI)	Unexplained	Top-model change (95% CI)
Subject	1.73	0.419 [0.363, 0.471]	58.1%	25.6% [19.5, 31.8]
Scene	1.73	0.418 [0.363, 0.470]	58.2%	19.5% [13.8, 25.1]
Motion	1.39	0.274 [0.222, 0.326]	72.6%	22.6% [16.9, 28.7]
Camera	1.37	0.266 [0.212, 0.320]	73.4%	26.7% [20.5, 32.8]

In Table 7, the other three factors explain only 26.6%–41.9% of held-out variance. Removing one factor changes the disjoint top-model set for 19.5%–26.7% of prompts. The temporary equal-rank aggregation used for this deletion test is a sensitivity probe, not a benchmark-wide diversity score.

Table 8: Factor-selective pair retrieval on VBench. The target response is in the top 10%, while all other factor responses are no higher than their medians.

Target factor	Prompts retrieved	Pair count
Subject	122/150 (81.3%)	286
Scene	112/150 (74.7%)	262
Motion	138/150 (92.0%)	484
Camera	140/150 (93.3%)	521

In Table 8, under the primary retrieval rule, a target-factor distance must be in the top 10% while all other factor distances remain below their medians within the same model–prompt group. Every factor retrieves such selective pairs for at least 74.7% prompts. This instance-level evidence complements the aggregate prediction tests by showing that each metric can expose changes not simultaneously flagged by the other three.

Together, the results support that the four metrics provide non-interchangeable diagnostic signals in the evaluated benchmark.

C Additional Results for the Main Benchmark Evaluation

This appendix reports the statistical details, ablations and sensitivity analyses omitted from Section 4.

C.1 Complete Model-Level Statistics

Prompt-paired significance analysis. To complement the model-level estimates in Table 9 and Table 10, we performed prompt-paired inference for the central rank reversal reported in the main text. The prompt was treated as the unit of analysis; each contrast retained prompts with valid scores for both models and used their prompt-wise MPD differences. We estimated 95% confidence intervals with 20,000 paired prompt-bootstrap resamples and applied two-sided Wilcoxon signed-rank tests, followed by Benjamini–Hochberg correction over all $6 \times 21 = 126$ dimension–model-pair comparisons.

Table 11 confirms the central cross-dimensional reversal. Relative to Hunyuan, Wan2.2-5B achieves higher MPD across all four content dimensions, with paired differences of 0.027–0.054 and confidence intervals that exclude zero ($q < 4.7 \times 10^{-8}$). Conversely, Hunyuan exceeds Wan2.2-5B by 0.107 in Motion and 0.394 in Camera (both $q < 5.0 \times 10^{-20}$). These prompt-paired results

Table 9: Complete model-level MPD statistics for the Semantic, Style and Subject dimensions. Each cell reports the mean ± prompt-level standard deviation and the 95% confidence interval obtained from 20,000 prompt-bootstrap resamples.

Model	Semantic	Style	Subject
Wan2.2-5B	0.169 ± 0.067 [0.160, 0.178]	0.234 ± 0.060 [0.226, 0.243]	0.447 ± 0.159 [0.425, 0.468]
Wan2.2-14B	0.144 ± 0.073 [0.134, 0.154]	0.224 ± 0.066 [0.215, 0.233]	0.412 ± 0.173 [0.388, 0.436]
CogVideo	0.142 ± 0.058 [0.134, 0.150]	0.214 ± 0.068 [0.205, 0.223]	0.414 ± 0.146 [0.394, 0.434]
Hunyuan	0.142 ± 0.060 [0.134, 0.150]	0.197 ± 0.063 [0.188, 0.205]	0.392 ± 0.163 [0.370, 0.414]
Wan2.7	0.158 ± 0.068 [0.149, 0.168]	0.201 ± 0.058 [0.193, 0.209]	0.397 ± 0.161 [0.375, 0.420]
HappyHorse	0.093 ± 0.054 [0.086, 0.101]	0.123 ± 0.042 [0.117, 0.129]	0.250 ± 0.141 [0.231, 0.269]
Seedance	0.160 ± 0.080 [0.150, 0.171]	0.162 ± 0.049 [0.156, 0.169]	0.380 ± 0.151 [0.359, 0.401]

Table 10: Complete model-level MPD statistics for the Scene, Motion, and Camera dimensions. Each cell reports the mean ± prompt-level standard deviation and the 95% confidence interval obtained from 20,000 prompt-bootstrap resamples.

Model	Scene	Motion	Camera
Wan2.2-5B	0.426 ± 0.140 [0.407, 0.446]	0.326 ± 0.104 [0.311, 0.340]	0.097 ± 0.151 [0.077, 0.118]
Wan2.2-14B	0.394 ± 0.154 [0.374, 0.416]	0.344 ± 0.109 [0.329, 0.359]	0.148 ± 0.235 [0.117, 0.181]
CogVideo	0.377 ± 0.124 [0.359, 0.394]	0.323 ± 0.089 [0.311, 0.336]	0.045 ± 0.119 [0.030, 0.063]
Hunyuan	0.378 ± 0.135 [0.360, 0.397]	0.433 ± 0.093 [0.420, 0.446]	0.490 ± 0.222 [0.458, 0.520]
Wan2.7	0.390 ± 0.130 [0.373, 0.408]	0.413 ± 0.090 [0.401, 0.425]	0.303 ± 0.225 [0.272, 0.334]
HappyHorse	0.223 ± 0.106 [0.208, 0.237]	0.326 ± 0.096 [0.313, 0.339]	0.193 ± 0.188 [0.168, 0.219]
Seedance	0.364 ± 0.132 [0.346, 0.382]	0.425 ± 0.091 [0.412, 0.437]	0.463 ± 0.236 [0.430, 0.495]

statistically substantiate the opposing content-versus-temporal diversity profiles reported in the main text. Together with the observation that six of the seven models span at least three rank positions across the six dimensions, they show that model diversity cannot be represented by a single global ordering and must instead be characterized dimension by dimension.

C.2 Real-Video Retrieval Diversity Reference

Video generation models are commonly trained on large-scale video–text datasets (Bain et al., 2021; Nan et al., 2025). To provide a reference for existing diversity metrics and estimate diversity exhibited by real web videos, we design a retrieval-based pipeline over approximately one million captions from OpenVid-1M (Nan et al., 2025). For each evaluation prompt, BGE-M3 (Chen et al., 2024) first retrieves the top 50 captions; we then rerank them by CLIP (Radford et al., 2021) video–text similarity computed over eight uniformly sampled frames, and evaluate the top 10 using the same frame-sampling and DiVid protocols as for generated videos.

As shown in Table 12, the retrieved videos achieve quality scores comparable to those of the evaluated generators and rank first in Semantic, Style, Subject, Scene, and six-dimension average diversity, but only fourth in Motion and Camera. This real-video reference further confirms that diversity is dimension-specific: even broad content variation does not imply equally rich temporal variation. It also suggests that drawing from large-scale web-video data alone may not provide sufficient supervision for motion and camera diversity, motivating dimension-aware data curation and structured temporal controls rather than global data rebalancing.

C.3 Quality–Faithfulness–Diversity Misalignment

In Table 13 and 14, across seven model means, the strongest negative association is IF–Subject ($\rho = -0.786$), but IF correlates positively with Motion and Camera. At the prompt level, 49 of 126 model-by-metric-by-dimension correlations remain significant after FDR correction: 35 are negative and 14 are positive.

Table 11: Prompt-paired validation of the central Wan2.2-5B–Hunyuan rank reversal. ΔMPD is positive for the first-named model. Confidence intervals are based on 20,000 paired prompt-bootstrap resamples; q values are obtained from two-sided Wilcoxon signed-rank tests with Benjamini–Hochberg correction across all 126 dimension–model-pair comparisons.

Dimension	Positive paired contrast	N	ΔMPD	95% bootstrap CI	Global BH q
Semantic	Wan2.2-5B − Hunyuan	206	0.027	[0.019, 0.035]	9.1×10^{-11}
Style	Wan2.2-5B − Hunyuan	206	0.038	[0.030, 0.046]	1.9×10^{-16}
Subject	Wan2.2-5B − Hunyuan	203	0.054	[0.037, 0.072]	1.2×10^{-8}
Scene	Wan2.2-5B − Hunyuan	203	0.047	[0.030, 0.063]	4.6×10^{-8}
Motion	Hunyuan − Wan2.2-5B	204	0.107	[0.089, 0.125]	5.0×10^{-20}
Camera	Hunyuan − Wan2.2-5B	204	0.394	[0.362, 0.424]	9.8×10^{-33}

Table 12: Quality and diversity of the retrieved real-video reference. Diversity ranks, shown in parentheses, are computed jointly with the seven evaluated generation models.

Source	AQ	IQ	TF	MS	Semantic	Style	Subject	Scene	Motion	Camera	Six-dim. Avg.
Real videos	0.574	0.707	0.980	0.991	0.239 (1)	0.261 (1)	0.523 (1)	0.540 (1)	0.407 (4)	0.243 (4)	0.369 (1)

C.4 Faithfulness–Diversity Decision

C.4.1 Threshold-Gated Diversity

For a threshold τ, only videos whose matched faithfulness is at least τ are eligible. To prevent models with more surviving videos from receiving more pairs, equal-count MPD samples exactly $k = 3$ or $k = 5$ faithful videos per model–prompt group and retains only prompts feasible for all seven models.

Style, Subject, and Scene are led by Wan2.2-5B throughout $\tau \in [0.50, 1.00]$, while Motion and Camera are led by Hunyuan. Semantic switches from Wan2.2-5B to Seedance at $\tau = 0.85$, but the margin is only 0.0015–0.0021 and Seedance's bootstrap top-1 frequency is 51.8%–55.2%; we treat this as a near tie rather than a robust reversal.

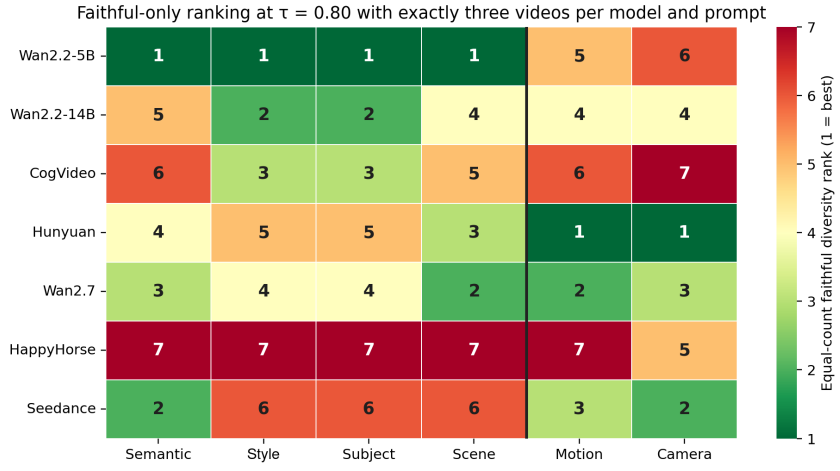

Figure 9: Equal-count diversity ranking at the primary setting $\tau = 0.8$, $k = 3$. Every cell displays the within-dimension model rank.

C.4.2 Continuous Preference Weights

Within each prompt and dimension, matched faithfulness and MPD are independently min–max normalized across the seven models to $\hat{F}, \hat{D} \in [0, 1]$. If one axis has zero range, it contributes no model distinction. Among Pareto-nondominated models, Balanced Pareto Win Rate selects the model

Table 13: Spearman correlations across seven model means.

Metric	Sem.	Sty.	Subj.	Scene	Mot.	Cam.
IF	-0.214	-0.679	-0.786	-0.429	0.393	0.500
AQ	0.179	-0.179	-0.393	-0.036	0.286	0.214
IQ	-0.071	-0.214	-0.357	0.000	0.000	0.071

Table 14: Prompt-level correlations significant after BH-FDR correction across 126 tests.

Metric	Negative	Positive	Total
IF	15	0	15
AQ	13	0	13
IQ	7	14	21
Total	35	14	49

closest to the ideal point using

$$S = 1 - \sqrt{(1 - w_D)(1 - \hat{F})^2 + w_D(1 - \hat{D})^2}. \tag{5}$$

The six-dimension macro curve averages win rates, not raw MPD, and is used only as a recommendation summary.

Wan2.7 ranks first for $w_D \in [0, 0.83]$, and Hunyuan ranks first for $w_D \in [0.84, 1]$; their top-2 coverage widths are 0.98 and 0.94. This recommendation is robust over broad preference ranges but remains dimension dependent, as shown in Figure 10.

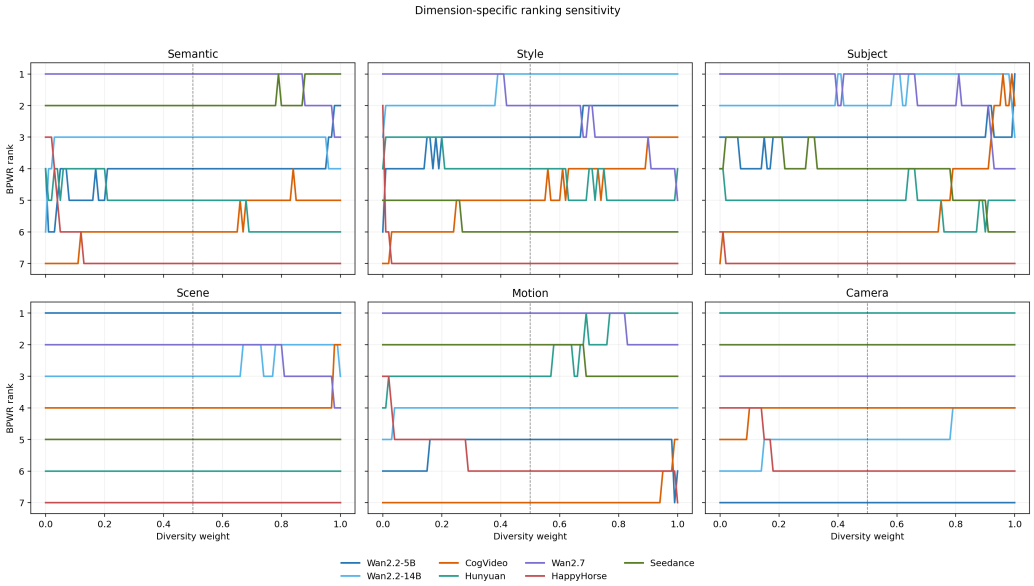

Figure 10: Dimension-specific BPWR ranks over the full diversity-weight range.

For a separate quality-oriented deployment scenario, we normalize AQ, IQ, and diversity within each prompt and assign weights 0.25/0.25/0.50. This is not the faithfulness–diversity score above: it answers which model balances perceptual quality and diversity when instruction fidelity is not the explicit constraint.

C.5 Cross-Prompt-Set Transfer

Tables 16 and 17 show the raw results on VBench and T2V-CompBench. Across the six diversity dimensions, 83.3% of the 21 model-pair orders are preserved on average. Style and Camera preserve

Table 15: Overall and factor-specific faithfulness.

Model	Overall IF	Subject	Scene	Motion	Camera
Wan2.2-5B	0.848	0.853	0.928	0.778	0.938
Wan2.2-14B	0.915	0.914	0.957	0.890	0.967
CogVideo	0.839	0.813	0.919	0.790	0.971
Hunyuan	0.893	0.898	0.937	0.842	0.977
Wan2.7	0.963	0.962	0.992	0.935	0.986
HappyHorse	**0.986**	**0.984**	**0.999**	**0.968**	**1.000**
Seedance	0.955	0.952	0.974	0.934	0.979

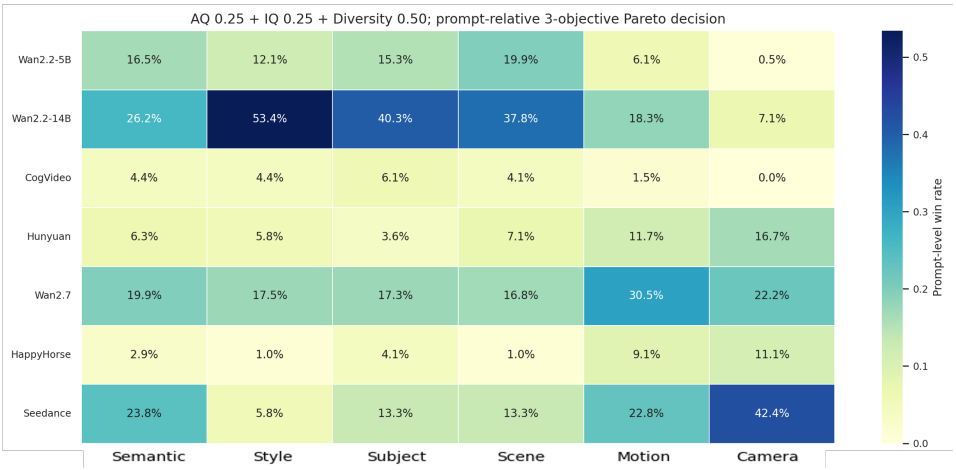

Figure 11: Prompt-level quality–diversity decision rates with AQ/IQ/Diversity weights 0.25/0.25/0.50.

all pairwise orders; Subject, Scene, and Motion show limited top-model changes under compositional prompts.

Table 16: Complete quality, faithfulness, and diversity profiles on the 150 VBench prompts.

	Faith / Quality					Prior Metrics		Holistic		Factor Level			
Model	IF	AQ	IQ	TF	MS	TCE	TIE	Semantic	Style	Subject	Scene	Motion	Camera
Wan2.2-5B	0.845	0.553	0.682	**0.982**	0.990	**5.555**	**13.422**	**0.184**	**0.241**	**0.473**	**0.450**	0.317	0.083
Wan2.2-14B	0.916	**0.620**	0.704	0.972	0.984	4.590	13.262	0.149	0.228	0.425	0.406	0.338	0.146
CogVideo	0.857	0.531	0.636	0.974	0.986	4.006	12.385	0.148	0.217	0.421	0.382	0.315	0.051
Hunyuan	0.899	0.549	0.670	0.965	**0.991**	5.318	13.263	0.154	0.206	0.416	0.399	**0.450**	**0.504**
Wan2.7	0.965	0.572	0.731	0.972	0.985	5.177	12.765	0.170	0.206	0.414	0.408	0.422	0.308
HappyHorse	**0.988**	0.605	**0.736**	0.978	0.991	2.806	11.487	0.099	0.129	0.272	0.238	0.337	0.202
Seedance	0.958	0.602	0.673	0.970	0.990	5.547	13.228	0.180	0.168	0.397	0.388	0.440	0.497

C.6 Cross-Model Complementarity

For each prompt and model pair, D_{within} averages the two matched within-model blocks, and D_{between} averages all cross-model pairs. We define descriptive complementarity as

$$\Delta_{\text{comp}} = D_{\text{between}} - D_{\text{within}}. \tag{6}$$

The union improves over the matched within baseline on 97.0% of prompt–dimension comparisons on average.

CogVideo and HappyHorse have relatively low standalone content diversity yet form the most complementary pair in four dimensions. Thus selecting two individually diverse models is not sufficient: routing should also consider whether their default output preferences differ.

Table 17: Complete quality, faithfulness, and diversity profiles on the 56 T2V-CompBench prompts.

	Faith / Quality					Prior Metrics		Holistic		Factor Level			
Model	IF	AQ	IQ	TF	MS	TCE	TIE	Semantic	Style	Subject	Scene	Motion	Camera
Wan2.2-5B	0.858	0.518	0.645	0.974	0.986	**4.564**	13.198	**0.131**	**0.216**	0.376	0.363	0.349	0.134
Wan2.2-14B	0.912	**0.598**	0.683	0.970	0.981	4.388	13.368	0.131	0.211	0.376	0.361	0.358	0.155
CogVideo	0.789	0.475	0.585	0.970	0.983	4.418	**13.407**	0.127	0.207	**0.395**	**0.363**	0.345	0.030
Hunyuan	0.876	0.517	0.641	0.965	0.991	4.351	12.546	0.109	0.173	0.329	0.323	0.389	**0.453**
Wan2.7	0.958	0.539	0.700	0.972	0.985	4.310	12.664	0.129	0.187	0.353	0.343	**0.390**	0.289
HappyHorse	**0.979**	0.566	**0.710**	**0.977**	**0.991**	2.679	11.342	0.078	0.107	0.190	0.181	0.298	0.169
Seedance	0.947	0.577	0.652	0.969	0.989	4.140	12.724	0.109	0.147	0.334	0.298	0.382	0.369

Table 18: Model pair with the largest mean complementarity in each dimension.

Dimension	Most complementary pair	Δ_{comp}
Semantic	CogVideo + HappyHorse	0.140
Style	CogVideo + HappyHorse	0.117
Subject	CogVideo + HappyHorse	0.230
Scene	CogVideo + HappyHorse	0.257
Motion	Wan2.2-5B + Seedance	0.139
Camera	CogVideo + Seedance	0.152

C.7 Dimension-Specific Case Studies

Global failures concern repeated concepts or appearance; composition failures concern subject or scene collapse; temporal failures concern repeated action or camera policies. Each shows four models selected to expose a clear MPD contrast.

Holistic. Semantic collapse appears when repeated samples preserve nearly the same concept realization, whereas Style collapse preserves color, viewpoint, texture, or layout even when the nominal subject changes. Figure 13 shows that the benchmark separates these two forms of global homogenization.

Composition exploration. Subject collapse retains nearly identical foreground identity or appearance, while Scene collapse retains the same environment despite repeated sampling. Figure 14 shows the cases.

Temporal exploration. Motion collapse repeats subject trajectories or deformation, while Camera collapse repeats the shot policy. Figure 15 shows the cases.

D Supplementary Materials for Section 5

This appendix supplements Section 5 with the controlled-prompt construction, statistical protocol, faithfulness diagnostics, and representative failure modes.

D.1 Complete Experimental Design and Prompt Construction

We use the fully specified prompt as the paired baseline and change only one target dimension at a time. The four settings are defined in Table 19. **Specified** provides the reference constraint, **Omitted** removes the target description, **Hinted** replaces it with a generic cue, and **Enumerated** provides five explicit candidates. Subject omission is undefined because a valid generation prompt must contain a subject. Each Enumerated candidate is a different prompt and is generated once.

Specified prompts were first drafted with a common template, then screened by three independent reviewers. A prompt was retained only when the four dimensions admitted multiple plausible alternatives, the target span could be replaced without changing the other spans, and the resulting sentence was grammatical and physically plausible. The controlled set contains 20 prompts covering people, animals, vehicles, indoor and outdoor environments, and urban and natural scenes.

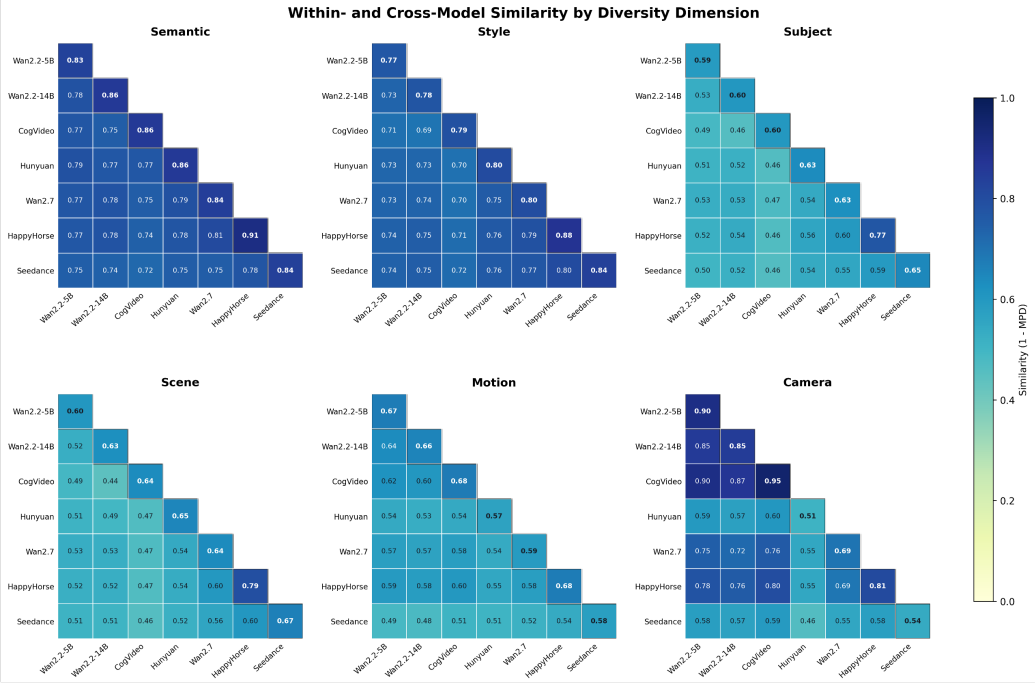

Figure 12: intra- and inter-model similarity matrices.

Table 19: Controlled prompt settings. Subject omission is undefined because a valid prompt requires a subject.

Setting	Construction	Diagnostic role
Specified	All four dimensions are explicitly specified.	Paired reference condition.
Omitted	Remove the target-dimension description.	Tests autonomous exploration.
Hinted	Replace the specific description with a generic cue.	Tests dimension activation.
Enumerated	Enumerate target candidates; generate one video per prompt.	Tests candidate realization.

For example, the golden-retriever prompt is constructed by replacing only one span at a time: the subject becomes `an animal`, the scene becomes `somewhere`, the motion becomes `in motion`, or the camera becomes `camera moves`. The complete example, including all five Enumerated prompts for each dimension, is given in Table 20. We use GPT-5 to generate Enumerated prompt. The complete prompts are in Figure 24, 25, 26, 27.

D.2 Statistical Details

For model m, semantic group i, target dimension d, and intervention r, we compute

$$\Delta^{(r)}_{m,i,d} = \mathrm{MPD}^{(r)}_{m,i,d} - \mathrm{MPD}^{(\mathrm{Specified})}_{m,i,d}. \tag{7}$$

Model means average this paired difference over the 20 groups. For the panel-level analysis, we first average the seven fixed models within each group and then use the 20 group means as the statistical units. We report a 20,000-sample group bootstrap confidence interval, block win rate, paired Hedges g_z, Wilcoxon signed-rank p values, and BH-adjusted q values. Full results are in Table 21 and 22.

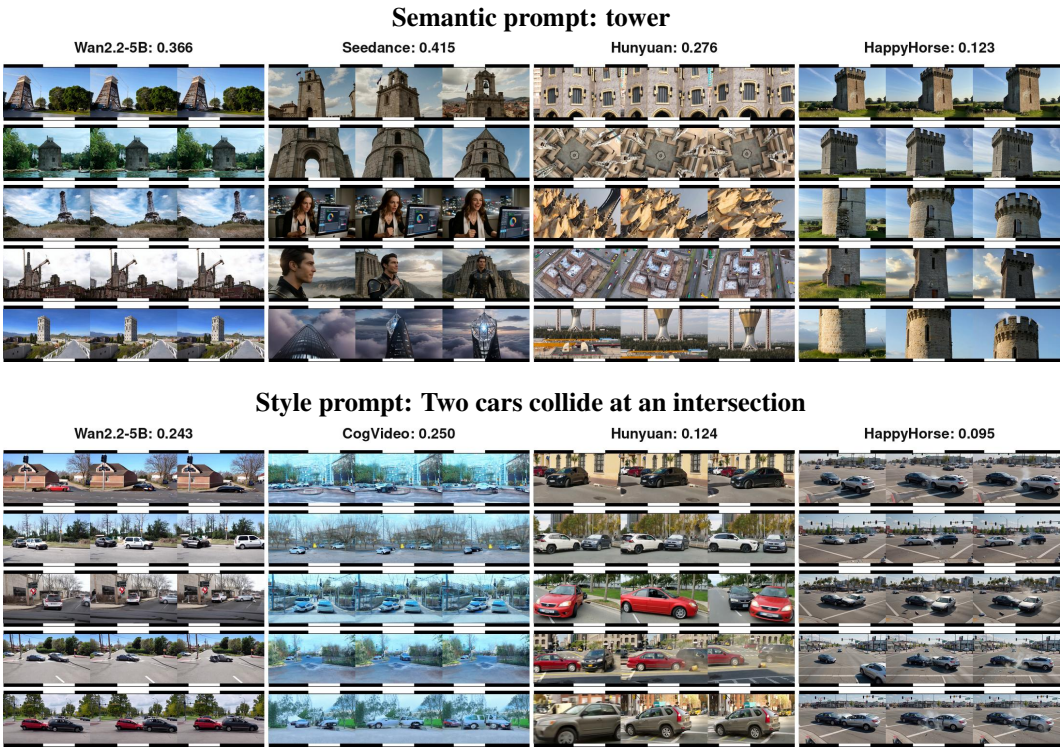

Figure 13: Holistic cases.



Figure 14: Composition cases.

Table 20: Complete controlled-prompt example for one semantic group. Only the target span changes within each intervention block.

Setting	Target	No.	Prompt
Specified	All	–	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, static camera
Hinted	Subject	–	An animal is resting on a quiet sandy beach at golden hour, static camera
Hinted	Scene	–	A large fluffy adult golden retriever is resting somewhere, static camera
Hinted	Motion	–	A large fluffy adult golden retriever is in motion on a quiet sandy beach at golden hour, static camera
Hinted	Camera	–	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, camera moves
Omitted	Scene	–	A large fluffy adult golden retriever is resting, static camera
Omitted	Motion	–	A large fluffy adult golden retriever is on a quiet sandy beach at golden hour, static camera
Omitted	Camera	–	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour
Enumerated	Subject	1	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, static camera
Enumerated	Subject	2	A small sleek black cat with green eyes is resting on a quiet sandy beach at golden hour, static camera
Enumerated	Subject	3	A tall spotted male giraffe with a long neck is resting on a quiet sandy beach at golden hour, static camera
Enumerated	Subject	4	A chubby gray baby seal with whiskers is resting on a quiet sandy beach at golden hour, static camera
Enumerated	Subject	5	A colorful adult macaw with bright feathers and a curved beak is resting on a quiet sandy beach at golden hour, static camera
Enumerated	Scene	1	A large fluffy adult golden retriever is resting in a cozy sunlit living room with wooden floors, static camera
Enumerated	Scene	2	A large fluffy adult golden retriever is resting on a quiet snowy field under an overcast sky, static camera
Enumerated	Scene	3	A large fluffy adult golden retriever is resting in a narrow neon-lit alley at night, static camera
Enumerated	Scene	4	A large fluffy adult golden retriever is resting on a grassy hilltop during a breezy sunrise, static camera
Enumerated	Scene	5	A large fluffy adult golden retriever is resting inside a rustic barn with warm lantern light, static camera
Enumerated	Motion	1	A large fluffy adult golden retriever is slowly wagging its tail on a quiet sandy beach at golden hour, static camera
Enumerated	Motion	2	A large fluffy adult golden retriever is trotting along the shoreline on a quiet sandy beach at golden hour, static camera
Enumerated	Motion	3	A large fluffy adult golden retriever is digging energetically in the sand on a quiet sandy beach at golden hour, static camera
Enumerated	Motion	4	A large fluffy adult golden retriever is stretching its front legs forward on a quiet sandy beach at golden hour, static camera
Enumerated	Motion	5	A large fluffy adult golden retriever is shaking water off its fur on a quiet sandy beach at golden hour, static camera
Enumerated	Camera	1	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, slow push-in camera
Enumerated	Camera	2	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, gentle pull-back camera
Enumerated	Camera	3	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, smooth left-to-right tracking camera
Enumerated	Camera	4	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, slow orbiting camera
Enumerated	Camera	5	A large fluffy adult golden retriever is resting on a quiet sandy beach at golden hour, rising crane camera

Table 21: Complete group-level uncertainty for the target responses. This table supplements the model-level means in the main paper.

Intervention	Dimension	Mean Δ [95% CI]	Positive blocks	g_z	p	q
Hinted	Subject	+0.197 [+0.144,+0.251]	19/20	1.49	1.91×10^{-6}	3.00×10^{-6}
Hinted	Scene	+0.056 [+0.032,+0.087]	17/20	0.81	3.15×10^{-5}	4.33×10^{-5}
Hinted	Motion	+0.069 [+0.044,+0.094]	18/20	1.13	1.31×10^{-4}	1.60×10^{-4}
Hinted	Camera	+0.210 [+0.184,+0.239]	20/20	3.05	9.54×10^{-7}	1.75×10^{-6}
Omitted	Scene	+0.051 [+0.025,+0.082]	18/20	0.73	1.97×10^{-4}	2.17×10^{-4}
Omitted	Motion	+0.011 [-0.007,+0.029]	12/20	0.26	0.139	0.139
Omitted	Camera	+0.092 [+0.069,+0.117]	20/20	1.58	9.54×10^{-7}	1.75×10^{-6}
Enumerated	Subject	+0.473 [+0.416,+0.528]	20/20	3.42	9.54×10^{-7}	1.75×10^{-6}
Enumerated	Scene	+0.370 [+0.304,+0.435]	20/20	2.31	9.54×10^{-7}	1.75×10^{-6}
Enumerated	Motion	+0.146 [+0.125,+0.170]	20/20	2.66	9.54×10^{-7}	1.75×10^{-6}
Enumerated	Camera	+0.321 [+0.288,+0.357]	20/20	3.85	9.54×10^{-7}	1.75×10^{-6}

D.3 Faithfulness Evaluation and Realization Cost

We compare Enumerated with Hinted within the same model, semantic group, and target dimension. Since Hinted prompts are less specific and therefore easier to satisfy, this contrast measures the realization cost of adding explicit control. Results are reported in Table 23.

Table 22: Adjacent-setting contrasts computed within each target dimension.

Contrast	Dimension	Mean difference [95% CI]	Positive blocks	Positive models	q
Hinted − Omitted	Scene	+0.005 [-0.007,+0.018]	10/20	4/7	0.298
Hinted − Omitted	Motion	+0.058 [+0.032,+0.087]	17/20	6/7	1.17×10^{-4}
Hinted − Omitted	Camera	+0.118 [+0.092,+0.147]	20/20	6/7	1.00×10^{-6}
Enumerated − Hinted	Subject	+0.275 [+0.250,+0.301]	20/20	7/7	1.00×10^{-6}
Enumerated − Hinted	Scene	+0.313 [+0.262,+0.364]	20/20	7/7	1.00×10^{-6}
Enumerated − Hinted	Motion	+0.077 [+0.057,+0.098]	20/20	7/7	1.00×10^{-6}
Enumerated − Hinted	Camera	+0.112 [+0.078,+0.143]	19/20	5/7	1.20×10^{-5}

Table 23: Faithfulness changes when moving from Hinted to explicit enumeration.

Dimension	Hinted mean	Enumerated mean	Enumerated − Hinted [95% CI]	Negative blocks	Negative models	q
Subject	4.593	4.497	-0.096 [-0.258,+0.048]	10/20	5/7	0.226
Scene	4.999	4.791	-0.204 [-0.254,-0.156]	19/20	7/7	8.90×10^{-5}
Motion	4.780	4.090	-0.690 [-0.887,-0.486]	18/20	7/7	3.82×10^{-5}
Camera	4.469	4.075	-0.393 [-0.541,-0.247]	18/20	6/7	8.90×10^{-5}

D.4 Failure Modes Revealed by Controlled Prompting

Prompt-gated default convergence. When a target dimension is replaced by a broad cue, the model may preserve a narrow default realization. In the Subject case (Figure 16), HappyHorse produces mostly the same dog under `an animal`: Specified, Hinted, and Enumerated MPD are 0.136, 0.147, and 0.871.

Dynamics. The same diagnostic appears in temporal dimensions (Figure 17). For Hunyuan on the road-person prompt, Motion MPD falls from 0.464 in Specified to 0.259 with a generic cue and 0.325 after omission, but rises to 0.613 under explicit action candidates. Camera shows the same pattern in the matched block: 0.225, 0.125, 0.216, and 0.752 for Specified, Hinted, Omitted, and Enumerated.

Realization-limited diversity. Explicit enumeration can enlarge the measured span while some candidates remain only partially realized. In the Hunyuan Motion case (Figure 18), Hinted videos all receive $F = 5$, whereas the five Enumerated candidates include scores 5, 5, 4, 5, and 3. In the corresponding Camera case, the Enumerated scores include one $F = 2$ despite a large MPD gain. These cases motivate reporting diversity and faithfulness separately.

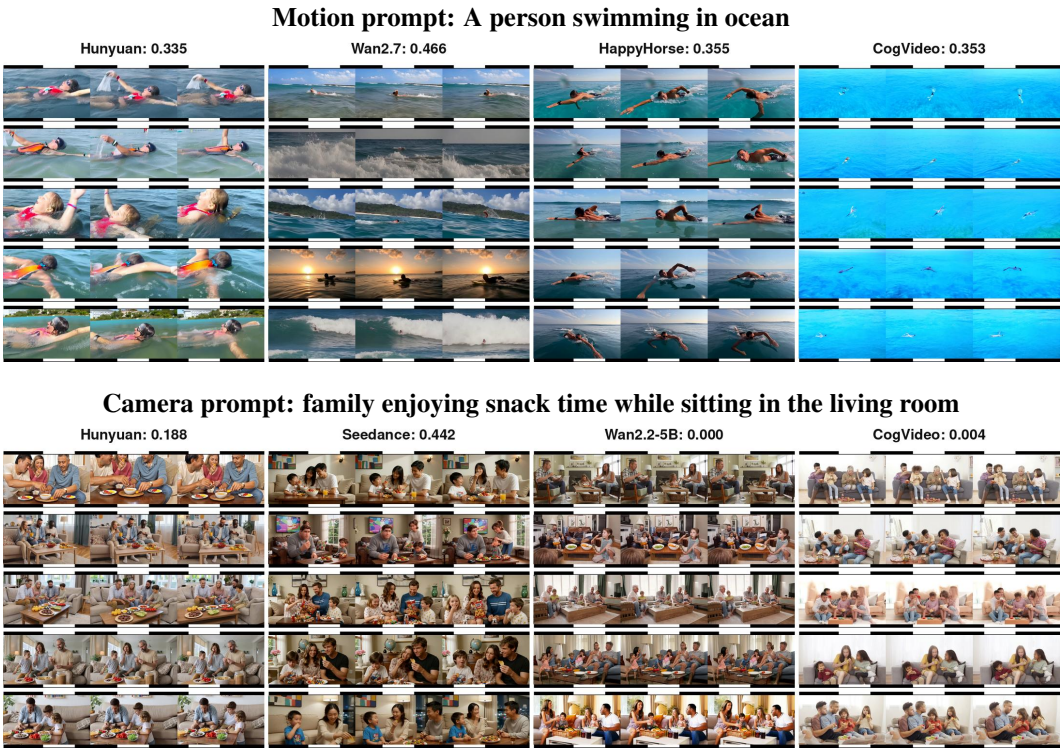

Figure 15: Temporal cases.

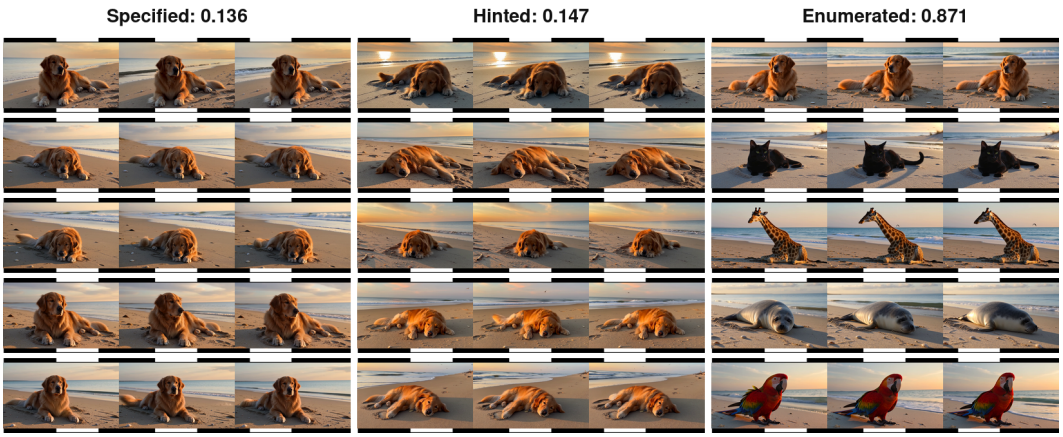

Figure 16: Prompt-gated default convergence.



Figure 17: Both cases use Hunyuan; the left column releases Motion and the right column releases Camera.

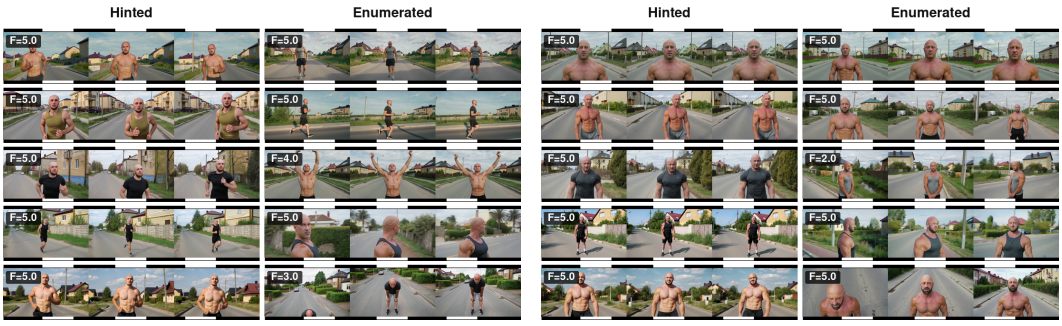

Figure 18: Realization-limited diversity. Left: Hunyuan Motion. Right: Hunyuan Camera. Each row displays the target-dimension Faithfulness score on the video.

Prompt for Subject Annotation

You are an annotation assistant for video-generation prompts. Your task is to annotate the subject field for each prompt.

Goal

subject refers to the **clearly visible, recognizable, foreground entity that can be reliably localized and segmented by models**.

Annotation rules

- **Use a single subject by default**, choosing the most central and clearly identifiable entity.
- **Only annotate multiple subjects when the prompt contains clearly coordinated core entities that should all visibly appear in the scene.**
  Example: a bird and a cat → ["bird", "cat"]
- **Use a generic category word, preferably a single word.**
  - golden retriever → dog
  - photographer → person
- **Nouns in prepositional phrases are usually not subjects**, unless they are clearly coordinated core entities. a cat on a table → ["cat"]
- **Objects of verbs are not automatically subjects**; include them only if they are also independent, salient, core entities.
  a person riding a bike → ["person"]
- Keep all original JSON fields unchanged and only add or correct subject.

Output format

- Return valid JSON
- Keep original fields
- subject must be a list of strings

Examples

- A blue car drives past a white picket fence → ["car"]
- A person is riding a bike → ["person"]
- a bird and a cat → ["bird", "cat"]
- Gwen Stacy reading a book → ["person"]

Please follow these rules strictly.

Figure 19: Prompt for Subject Annotation.

Prompt for Evaluating Subject Faithfulness

You are a professional AI-generated video evaluation expert. You will be given a text Prompt used to generate a video, along with the corresponding generated video.
Your task: Evaluate the generated video's adherence/faithfulness to the text Prompt from only the [Subject Fidelity] dimension.

Text Prompt: "{prompt_text}"

Dimension Definition - "Who/What" and their static attributes
Evaluation scope (includes):
- Whether the subjects/objects specified in the prompt exist
- Whether the number of subjects is correct
- Inherent attributes of subjects (color, shape, material, clothing, species, identity features, etc.)
- Static positional relationships between multiple subjects

Evaluation scope (excludes):
- Any actions/motion processes (not considered in this dimension)
- Background elements that do not belong to the subjects themselves (not considered in this dimension)
- Camera movement (not considered in this dimension)

Scoring Criteria
- 5 = Fully matches Prompt requirements, no deviation
- 4 = Minor deviations exist, but do not affect overall understanding
- 3 = Partially matches, some elements described in the Prompt are missing or incorrect
- 2 = Clearly inconsistent, most elements are incorrect
- 1 = Completely irrelevant / contradicts the Prompt
- "N/A" = Only when the Prompt has absolutely no relevant description or requirements for this dimension

Additional Rules
- If the video contains extra subjects/objects not mentioned in the Prompt (model hallucination), points should be deducted in this dimension.
- Only evaluate content related to subjects; do not incorporate issues related to actions, scenes, or camera into this dimension.

Output Format
Do not output thinking or analysis process. Directly output the following format:

reason: <Brief analysis of subject fidelity adherence> <score>: <1-5 or "N/A">

Figure 20: Prompt for Evaluating Subject Faithfulness.

Prompt for Evaluating Scene Faithfulness

You are a professional AI-generated video evaluation expert. You will be given a text Prompt used to generate a video, along with the corresponding generated video.
Your task: Evaluate the generated video's adherence/faithfulness to the text Prompt from only the [Scene/Environment] dimension.

Text Prompt: "{prompt_text}"

Dimension Definition - "Where" and overall atmosphere/style setting
Evaluation scope (includes):
- Background/location/scene settings specified in the prompt
- Time (daytime/nighttime)
- Weather
- Lighting conditions
- Overall visual style (e.g., "cyberpunk style", "black-and-white film feel", "watercolor animation style", etc., if specified in the Prompt)
- Overall atmosphere/mood

Evaluation scope (excludes):
- Subjects themselves and their attributes (not considered in this dimension)
- Any subject movement (not considered in this dimension)
- Camera movement (not considered in this dimension)

Scoring Criteria
- 5 = Fully matches Prompt requirements, no deviation
- 4 = Minor deviations exist, but do not affect overall understanding
- 3 = Partially matches, some elements described in the Prompt are missing or incorrect
- 2 = Clearly inconsistent, most elements are incorrect
- 1 = Completely irrelevant / contradicts the Prompt
- "N/A" = Only when the Prompt has absolutely no relevant description or requirements for this dimension

Additional Rules
- If the video contains extra scene elements not mentioned in the Prompt (model hallucination), points should be deducted in this dimension.
- Only evaluate content related to scene/environment; do not incorporate issues related to subject attributes, actions, or camera into this dimension.

Output Format
Do not output thinking or analysis process. Directly output the following format:

reason: <Brief analysis of scene/environment adherence> <score>: <1-5 or "N/A">

Figure 21: Prompt for Evaluating Scene Faithfulness.

Prompt for Evaluating Motion Faithfulness

You are a professional AI-generated video evaluation expert. You will be given a text Prompt used to generate a video, along with the corresponding generated video.
Your task: Evaluate the generated video's adherence/faithfulness to the text Prompt from only the [Action/Motion Accuracy] dimension.

Text Prompt: "{prompt_text}"

Dimension Definition - "What is happening", i.e., dynamic behavior
Evaluation scope (includes):
- Actions and motion processes described in the prompt
- Interactions between subjects
- Direction/speed/manner of movement
- Sequential order of multiple actions
- Movement trajectories or posture changes of subjects themselves

Evaluation scope (excludes):
- Camera movement itself (not considered in this dimension)
- Static positional relationships (not considered in this dimension)
- Inherent attributes of subjects (not considered in this dimension)
- Background/scene elements (not considered in this dimension)

Scoring Criteria
- 5 = Fully matches Prompt requirements, no deviation
- 4 = Minor deviations exist, but do not affect overall understanding
- 3 = Partially matches, some elements described in the Prompt are missing or incorrect
- 2 = Clearly inconsistent, most elements are incorrect
- 1 = Completely irrelevant / contradicts the Prompt
- "N/A" = Only when the Prompt has absolutely no relevant description or requirements for this dimension

Additional Rules
- If the video contains extra actions not mentioned in the Prompt (model hallucination), points should be deducted in this dimension.
- Only evaluate content related to actions/motion; do not incorporate issues related to subject attributes, scenes, or camera into this dimension.

Output Format
Do not output thinking or analysis process. Directly output the following format:

reason: <Brief analysis of action/motion accuracy adherence> <score>: <1-5 or "N/A">

Figure 22: Prompt for Evaluating Motion Faithfulness.

Prompt for Evaluating Camera Faithfulness

You are a professional AI-generated video evaluation expert. You will be given a text Prompt used to generate a video, along with the corresponding generated video.
Your task: Evaluate the generated video's adherence/faithfulness to the text Prompt from only the [Camera/Camera Movement] dimension.

Text Prompt: "{prompt_text}"

Dimension Definition - "How it is shot", referring only to camera perspective-related content
Evaluation scope (includes):
- Camera movement methods specified in the prompt (push in, pull out, pan, track, drone aerial shot, follow shot, etc.)
- Camera angle (overhead shot, low-angle shot, eye-level, etc., if specified in the Prompt)
- Shot size (close-up, medium shot, wide shot, etc., if specified in the Prompt)

Evaluation scope (excludes):
- Movement of subjects themselves (not considered in this dimension)
- If the Prompt is describing the subject's own action and there is no explicit camera context, it should not be classified under this dimension
- Subject attributes (not considered in this dimension)
- Scene/background (not considered in this dimension)

Scoring Criteria
- 5 = Fully matches Prompt requirements, no deviation
- 4 = Minor deviations exist, but do not affect overall understanding
- 3 = Partially matches, some elements described in the Prompt are missing or incorrect
- 2 = Clearly inconsistent, most elements are incorrect
- 1 = Completely irrelevant / contradicts the Prompt
- "N/A" = Only when the Prompt has absolutely no relevant description or requirements for this dimension

Additional Rules
- Only evaluate content related to camera/camera movement; do not incorporate issues related to subject attributes, actions, or scenes into this dimension.
- Pay attention to distinguishing between "the subject is moving" and "the camera is moving"; only the latter belongs to this dimension.

Output Format
Do not output thinking or analysis process. Directly output the following format:

reason: <Brief analysis of camera/camera movement adherence> <score>: <1-5 or "N/A">

Figure 23: Prompt for Evaluating Camera Faithfulness.

Prompt for Subject Rewriting

You are a video generation prompt rewriting expert. Your task is to concretize the generic subject in a given prompt by replacing the vague subject description with specific, rich subject descriptions.

Constraints:
1. Only modify the subject description portion; all other dimensions (action, scene, camera movement) must remain completely unchanged
2. Variants should be maximally semantically orthogonal (maximally diverse)
3. Each variant should be reasonable and natural—something a user would plausibly write—with smooth, coherent phrasing
4. Maintain uniform formatting and similar length across variants
5. Each subject description should include 2-4 visual characteristics

Prompt to rewrite:

Generate k diverse rewritten prompts. You may only modify the subject portion; all other parts must remain completely unchanged.

Figure 24: Prompt for Subject Rewriting.

Prompt for Scene Rewriting

You are a video generation prompt rewriting expert. Your task is to concretize the generic scene in a given prompt by replacing the vague scene description "somewhere" with specific, rich scene descriptions.

Constraints:
1. Only modify the scene description portion "somewhere"; all other dimensions (subject, action, camera movement) must remain completely unchanged
2. Variants should be maximally semantically orthogonal (maximally diverse), covering different environment types (indoor/outdoor, natural/man-made, different times of day, different weather conditions, different cultural styles, etc.)
3. Each variant should be reasonable and natural—something a user would plausibly write—with smooth, coherent phrasing
4. Maintain uniform formatting and similar length across variants
5. Each scene description should include 2-4 environmental characteristics

Prompt to rewrite:

Concretize "somewhere" and generate k diverse rewritten prompts. You may only modify the scene portion "somewhere"; all other parts must remain completely unchanged.

Figure 25: Prompt for Scene Rewriting.

Prompt for Motion Rewriting

You are a video generation prompt rewriting expert. Your task is to concretize the generic action in a given prompt by replacing the vague action description "is in motion" with specific, rich action descriptions.

Constraints:
1. Only modify the action description portion "is in motion"; all other dimensions (subject, scene, camera movement) must remain completely unchanged
2. Variants should be maximally semantically orthogonal (maximally diverse), covering different ranges of motion, different body parts, or different action types
3. Each variant should be reasonable and natural—something a user would plausibly write—with smooth, coherent phrasing
4. Maintain uniform formatting and similar length across variants
5. Each action description should be specific and visually concrete, incorporating different manners of movement or speed modifiers

Prompt to rewrite:

Concretize "is in motion" and generate k diverse rewritten prompts. You may only modify the action portion; all other parts must remain completely unchanged.

Figure 26: Prompt for Motion Rewriting.

Prompt for Camera Rewriting

You are a video generation prompt rewriting expert. Your task is to concretize the generic camera movement in a given prompt by replacing the vague camera movement description "camera moves" with specific, rich camera movement descriptions.

Constraints:
1. Only modify the camera movement description portion "camera moves"; all other dimensions (subject, action, scene) must remain completely unchanged
2. Variants should be maximally semantically orthogonal (maximally diverse), covering different camera movement types
3. Each variant should be reasonable and natural—something a user would plausibly write—with smooth, coherent phrasing
4. Maintain uniform formatting and similar length across variants
5. Each camera movement description should clearly specify how the camera moves, incorporating different directions or speeds

Prompt to rewrite:

Concretize "camera moves" and generate k diverse rewritten prompts. You may only modify the camera movement portion "camera moves"; all other parts must remain completely unchanged.

Figure 27: Prompt for Camera Rewriting.