DIAL: A Benchmark for Evaluating Contrastive Instruction Isolation in Multi-Shot Video Generation

Dongwoo Lee

February 2026

Abstract

As text-to-video (T2V) generation evolves toward sophisticated multi-shot narratives (T2MSV), current evaluation paradigms remain severely constrained by single-shot assumptions. Traditional metrics, such as Fréchet Video Distance (FVD) and frame-wise CLIP similarity, inherently fail to penalize **"Context Bleeding"**—a phenomenon where models conflate distinct temporal instructions—and inadvertently reward the **"Static Video Trap,"** where models achieve artificially high temporal consistency by redundantly copying content across shots, completely ignoring narrative transitions. In this paper, we introduce **DIAL** (Diagonal Instruction ALignment), a comprehensive evaluation benchmark structured into two orthogonal tracks: **Track S (Semantic Leap)** for drastic environmental shifts, and **Track M (Motion Continuity)** for fine-grained spatial camera control. Central to our benchmark is **Diagonal Semantic Alignment (DSA)**. By leveraging a column-wise softmax normalization, DSA shifts the paradigm from absolute alignment to **Contrastive Instruction Isolation**, explicitly enforcing a zero-sum game that assigns a strict zero score to static, instruction-agnostic generations. Through an exhaustive evaluation of 9 state-of-the-art models across 1,000 stress-test scenarios, we uncover a fundamental **"Double-Kill" dilemma**: current monolithic architectures either preserve identity at the expense of dynamics (the Static Trap) or prioritize dynamics but suffer from severe **Identity Amnesia**. Our findings establish DIAL as a critical diagnostic foundation for developing next-generation decoupled video priors.

1 Introduction

The field of generative video AI is rapidly transitioning from synthesizing blurry, single-shot clips to attempting complex, cinematic multi-shot narratives, a paradigm shift we define as **Text-to-Multi-Shot Video (T2MSV)**. This evolution promises to revolutionize digital storytelling, enabling the creation of coherent films and dynamic visual narratives from simple textual sequences. However,

as the complexity of the generated content increases, our ability to evaluate it has stagnated. Current evaluation methodologies remain deeply anchored in a single-shot mindset, relying on metrics that fail to account for the unique spatiotemporal challenges of multi-shot synthesis.

Existing quality and consistency metrics, such as Fréchet Video Distance (FVD) [**?**] or frame-wise CLIP similarity [**?**], evaluate generated videos holistically. In the context of T2MSV, this holistic approach introduces a critical structural vulnerability: it inadvertently rewards generative models that produce entirely static, unchanging videos, even when those models are explicitly prompted to execute highly dynamic scene transitions. We define this pervasive failure phenomenon as the **Static Video Trap**. For example, if a model is prompted with a sequence demanding a narrative leap—such as jumping from a "dense rainforest" to "deep outer space"—a model that stubbornly generates a continuous rainforest scene will score artificially high on traditional temporal consistency metrics. This creates the **Global Similarity Paradox**, where conventional metrics inherently reward the failure to transition, thereby masking a fundamental inability to follow complex shot-level semantic instructions.

Furthermore, current evaluation paradigms are blind to **Context Bleeding**, where semantic elements from one prompt erroneously persist into subsequent shots. This failure is often obscured by the use of absolute CLIP scores, which measure overall alignment but fail to quantify the *isolation* of instructions. Ironically, we observe that naive baselines—such as concatenating independently generated shots—can achieve higher text alignment scores than sophisticated end-to-end models, a paradox that further underscores the need for a contrastive evaluation framework. When models do attempt to escape the Static Trap by executing dynamic transitions, they frequently succumb to **Identity Amnesia**, wherein the model loses the fine-grained visual features of the main subject across shots. This suggests a fundamental **"Double-Kill" dilemma** in current monolithic architectures: they can either preserve identity at the expense of dynamics or prioritize dynamics at the expense of identity, but they struggle to achieve both.

In this work, we argue that T2MSV requires more than mere alignment; it demands **Temporal Instruction Isolation**—the ability to execute a specific prompt exclusively at a specific time while actively rejecting adjacent contextual noise. To address this, we introduce **DIAL** (Diagonal Instruction ALignment), a comprehensive evaluation benchmark designed to diagnose these exact capabilities through a 4D decoupled evaluation framework. Our benchmark categorizes evaluation into two orthogonal tracks: **Track S (Semantic Leap)** tests narrative diversity by requiring radical environment changes while strictly preserving the subject, and **Track M (Motion Continuity)** evaluates spatial integrity, where the background must remain consistent while the camera executes specific physical motions.

To quantitatively support this taxonomy, we propose a 4D decoupled evaluation pipeline assessing subject identity, background dynamics, cut sharpness, and instruction alignment. Central to this pipeline is **Diagonal Semantic Alignment (DSA)**, a novel metric that applies column-wise softmax normal-

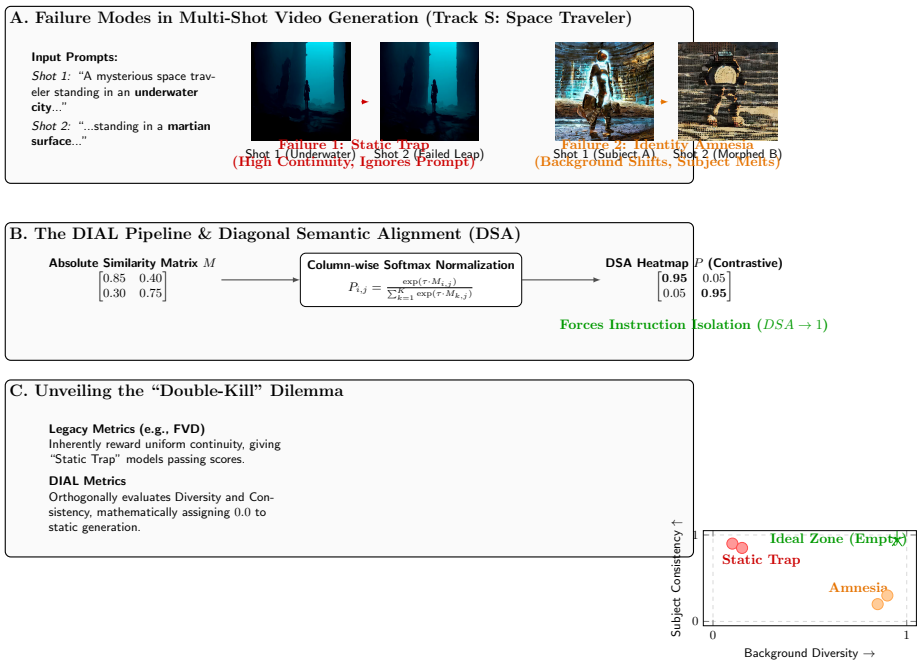


Figure 1: Overview of the **DIAL** Benchmark and the **DSA** Pipeline. **(A)** We utilize actual generated frames from our stress-test scenarios (e.g., Track S) to demonstrate how current models exploit traditional metrics via the **Static Trap** or fail visually via **Identity Amnesia**. **(B)** Our **Diagonal Semantic Alignment (DSA)** leverages a column-wise softmax to mathematically guarantee instruction isolation and explicitly penalize context bleeding. **(C)** This rigorous formulation successfully uncovers the underlying "Double-Kill" structural dilemma, proving that current architectures fail to balance spatiotemporal dynamics.

ization to shot-prompt similarity matrices. This formulation forces a zero-sum game, demanding exclusivity and penalizing context bleeding, while mathematically guaranteeing a zero score for static, instruction-ignoring videos.

Our core contributions are summarized as follows:

- We systematize the critical failure modes of T2MSV: the **Static Trap**, **Context Bleeding**, and **Identity Amnesia**, highlighting how legacy metrics inadvertently reward these failures.
- We propose **DIAL**, a benchmark featuring 1,000 stress-test scenarios, alongside a 4D decoupled evaluation pipeline.
- We introduce **Diagonal Semantic Alignment (DSA)**, a mathematically rigorous metric that utilizes contrastive normalization to penalize context bleeding and ensure a zero score for static generation.

- We conduct an exhaustive evaluation of 9 state-of-the-art models, uncovering the **"Double-Kill" Pareto frontier** and establishing a diagnostic foundation for next-generation decoupled video priors.

2 Related Work

2.1 Text-to-Video Generation Architectures

The rapid evolution of diffusion models has bifurcated the field of video generation into two distinct architectural paradigms: **Foundation T2V Models** and **Specialized T2MSV Frameworks**. Foundation models, such as CogVideoX [**?**], LTX-Video [**?**], and Stable Video Diffusion (SVD) [**?**], scale up massive video datasets to learn raw spatiotemporal priors. While these models excel at rendering high-fidelity, single-shot motions, they critically lack the explicit control mechanisms required for complex, multi-shot narrative sequencing. To address this limitation, Specialized T2MSV Frameworks aim to guide these base models toward narrative generation via complex inference pipelines. Methods such as StoryDiffusion [**?**] leverage consistent self-attention modules to enforce character tracking across frames, whereas frameworks like DirecT2V [**?**] and Mora [**?**] utilize multi-agent systems to explicitly enforce shot boundaries and camera directions. Alternatively, tuning-free inference methods such as FreeNoise [**?**] and AnimateDiff [**?**] attempt to extend video length or animate personalized models. Despite these sophisticated architectural and inference-time interventions, our empirical evaluation reveals that both paradigms still suffer profoundly from unintended prompt bleeding and catastrophic morphing artifacts during multi-shot generation.

2.2 The Flaws of Current T2MSV Evaluation

Current evaluation suites for video generation, such as VBench [**?**], rely on a taxonomy of metrics that are fundamentally misaligned with the cinematic grammar required by T2MSV, leading to systemic evaluation blind spots.

The Middle-Frame Fallacy. Many current studies evaluate identity and background consistency by sampling merely a single middle frame per shot. This approach is critically flawed as it fails to penalize temporal artifacts like morphing, where objects unnaturally melt into one another instead of cutting cleanly. Our framework overcomes this by employing a 4D decoupled evaluation pipeline that assesses consistency across all generated frames.

The Uniform Continuity Bias. Traditional temporal coherence metrics assume that lower frame-wise distance (e.g., LPIPS [**?**]) is always better. This bias unfairly penalizes intentional, cinematic cuts, treating sharp transitions as visual artifacts. We resolve this by introducing *Cut-Transition Sharpness*, which demands high perceptual divergence exclusively when a shot transition is prompted, effectively validating the model's ability to execute a clean break.

Absolute vs. Contrastive Alignment. Existing metrics rely on absolute

CLIP [**?**] scores to measure text-video alignment. However, absolute scores are blind to **Context Bleeding**—where a model hallucinates elements from previous or future prompts into the current shot. Ironically, naive baselines that concatenate independent shots often achieve higher absolute alignment than end-to-end models. Our proposed **Diagonal Semantic Alignment (DSA)** shifts the paradigm to contrastive instruction isolation. By leveraging a column-wise softmax normalization, DSA forces a zero-sum game that explicitly demands exclusivity and penalizes context bleeding, ensuring that instructions are isolated and executed only when intended.

The Global Similarity Paradox. Evaluating character consistency using holistic image similarity leads to a contradictory assessment: if a model correctly executes a drastic background shift as instructed, its holistic structural similarity artificially drops, penalizing the model for strictly following semantic instructions. We resolve this by decoupling the foreground subject from the background environment via precise DINOv2 [**?**] masking, allowing for a more granular and accurate diagnosis of generative successes and failures.

3 The DIAL Benchmark and 4D Methodology

We propose the **DIAL** (Diagonal Instruction ALignment) benchmark, a meticulously controlled stress-test dataset designed specifically to expose the structural blind spots of current video generation models. Unlike generic quality benchmarks that rely on massive, uncurated prompt pools, our benchmark consists of 1,000 highly targeted multi-shot scenarios. This scale is optimal for exposing the "Double-Kill" dilemma while remaining computationally feasible for our rigorous 4D evaluation pipeline. Table 1 details our comprehensive taxonomy, dividing the 1,000 scenarios equally into Track S and Track M.

Table 1: The Comprehensive Taxonomy of the **DIAL** Benchmark (1,000 Scenarios). Each track features 500 prompts distributed across 4 sub-categories to test extreme edge cases, preventing models from exploiting uniform continuity.

Track	Sub-category	Count	Objective	Target Vulnerability
Track S (Leap)	Spatial	125	Shift to entirely different physical locations (e.g., Jungle → Space)	Identity Amnesia
	Temporal	125	Shift across different eras (e.g., Medieval → Cyberpunk)	Identity Amnesia
	Atmosphere	125	Extreme weather/environmental changes in the same location	Static Trap
	Scale	125	Microscopic to macroscopic perspective shifts	Identity Amnesia
Track M (Motion)	Translational	125	X/Y axis camera panning without background breakdown	Static Trap
	Depth	125	Z axis zoom in/out maintaining subject identity	Identity Amnesia
	Tracking/Orbit	125	Complex 360-degree rotational rendering	Background Hallucination
	Compound	125	Mixed instructions (e.g., Pan Right + Zoom In)	Context Bleeding

3.1 Evaluation Scenarios: Track S and Track M

As illustrated in Figure 2, our evaluation framework categorizes generation scenarios into two conceptually orthogonal extremes to stress-test narrative robustness and spatial coherence.

Track S (Semantic Leap): This track evaluates narrative diversity by requiring radical environmental shifts while strictly preserving subject identity. This

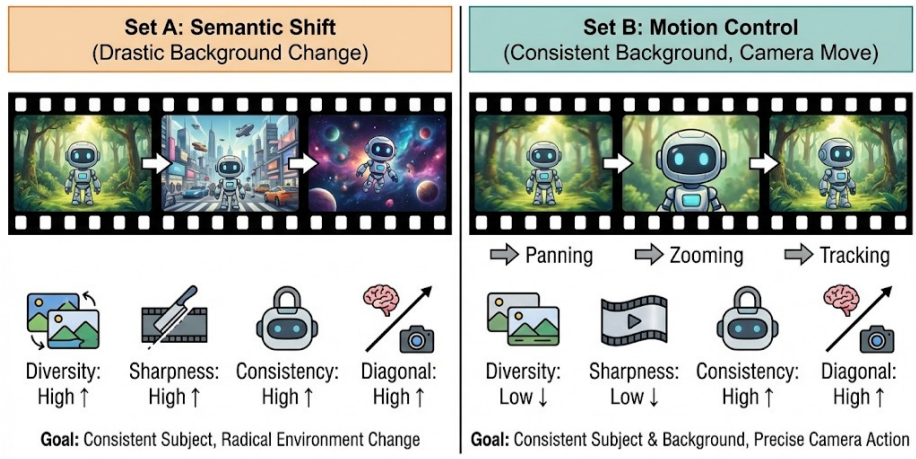


Figure 2: Conceptual comparison of our two evaluation tracks. (Left) **Track S: Semantic Leap** tests narrative diversity, requiring radical environment changes while preserving the subject. (Right) **Track M: Motion Continuity** evaluates spatial integrity, where the background must remain consistent while the camera executes specific motions.

is specifically designed to expose models that fail to execute semantic leaps, falling into the **Static Trap**.

Track M (Motion Continuity): In contrast, Track M evaluates spatial integrity, where the background must remain perfectly consistent while the camera executes specific physical motions (e.g., panning or zooming). This track tests the model's ability to maintain a consistent world state under dynamic camera control.

3.2 Formulation of 4D Decoupled Metrics

To diagnose generative failures accurately, we introduce four strictly decoupled metrics designed with robust mathematical penalty mechanisms.

1. Subject Consistency ($\mathcal{C}_{subj}$). We measure the semantic identity preservation of the core subject using DINOv2 embeddings. The subjects are localized using Grounding DINO to prevent background bias. Crucially, we enforce a strict **Error Propagation Penalty**: if the masking model fails to detect the subject due to severe identity amnesia or catastrophic morphing into the background, the frame is immediately assigned a hard consistency score of 0.0. This ensures that catastrophic generation failures are actively penalized rather than ignored.

2. Background Diversity ($\mathcal{D}_{bg}$). To formally identify the *Static Trap*, we measure the variance of the background environment embeddings. $\mathcal{D}_{bg}$ is calculated as the mean perceptual distance from the average background embedding across shots. A low score in Track S definitively confirms the model has failed

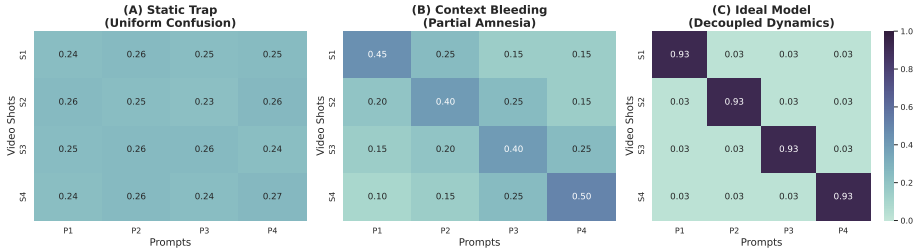


Figure 3: Diagonal Semantic Alignment (DSA) Heatmaps. (A) Models in the **Static Trap** exhibit a uniform probability distribution across prompts. (B) Models suffering from **Context Bleeding** exhibit noisy diagonal alignment. (C) An **Ideal Decoupled Model** achieves sharp, precise diagonal alignment, proving independent shot execution.

to execute the semantic leap.

3. Cut-Transition Sharpness ($\mathcal{S}_{cut}$). Traditional metrics assume a fixed cut time, which is unrealistic. To solve this, we introduce a **Dynamic Sliding Window Peak Prominence** algorithm. We calculate the LPIPS distance across all adjacent frames within a window W and define sharpness as the peak distance normalized by the surrounding mean:

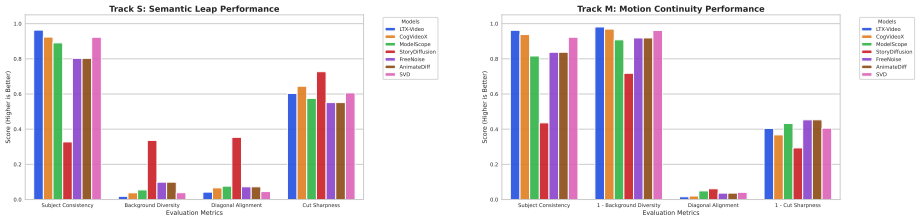
$$\mathcal{S}_{cut} = 1 - \frac{1}{\frac{\max_{t \in W} \text{LPIPS}(f_t, f_{t+1})}{\text{mean}(\text{LPIPS}) + \epsilon} + 1}$$

This mathematically differentiates a sharp, deliberate cinematic cut from a slow, blurry fade or morphing artifact.

4. Diagonal Semantic Alignment (DSA). Traditional CLIP scores measure absolute alignment, which is blind to context bleeding. We solve this by applying a column-wise softmax to the $K \times K$ shot-prompt similarity matrix M:

$$P_{i,j} = \frac{\exp(\tau \cdot M_{i,j})}{\sum_{k=1}^{K} \exp(\tau \cdot M_{k,j})}, \quad DSA = \max\left(0, \frac{\frac{1}{K}\sum_{i=1}^{K} P_{i,i} - \frac{1}{K}}{1 - \frac{1}{K}}\right)$$

Hyperparameter Robustness: The temperature τ is not an arbitrary hyperparameter, but the learned logit scale inherent to the pre-trained CLIP model (typically ~ 100.0). As detailed in our Appendix Ablation Studies, DSA consistently exposes the Static Trap ($DSA \approx 0$) regardless of the scale, proving it is a structurally robust evaluator of multi-shot independence. By leveraging a column-wise softmax, DSA shifts the paradigm from absolute alignment to **Contrastive Instruction Isolation**, forcing a zero-sum game that assigning a strict zero score to static, instruction-agnostic generations.



(a) Track S: Semantic Leap (b) Track M: Motion Continuity

Figure 4: Grouped Bar Charts of Model Performance. In **Track S**, models must achieve high scores across all metrics. In **Track M**, we plot inverted axes for Diversity and Sharpness $(1 - x)$ such that higher bars consistently represent better performance across all scenarios.

Table 2: Main Results on **Track S (Semantic Leap)** and **Track M (Motion Continuity)**. Track S demands narrative-driven dynamics (↑ is better for all). Track M demands spatial stability (↓ is better for Diversity and Sharpness). Best in **bold**.

		Track S (Semantic Leap)				Track M (Motion Continuity)			
Category	**Method**	**Cons. ↑**	**Div. ↑**	**DSA ↑**	**Sharp. ↑**	**Cons. ↑**	**Div. ↓**	**DSA ↑**	**Sharp. ↓**
Foundation	LTX-Video [?]	**0.963**	0.018	0.042	0.603	**0.962**	**0.019**	0.016	0.596
Foundation	CogVideoX [?]	0.924	0.038	0.066	0.644	0.938	0.031	0.020	0.632
Foundation	SVD [?]	0.923	0.039	0.045	0.606	0.923	0.039	0.041	0.594
Foundation	ModelScope	0.891	0.055	0.076	0.575	0.816	0.092	0.049	0.568
Framework	AnimateDiff [?]	0.802	0.099	0.072	0.551	0.837	0.081	0.036	**0.547**
Framework	FreeNoise [?]	0.802	0.099	0.072	0.551	0.837	0.081	0.036	**0.547**
Framework	StoryDiffusion [?]	0.328	**0.336**	**0.354**	**0.727**	0.436	0.282	**0.061**	0.707

4 Experiments

Our evaluation on the **DIAL** benchmark reveals systematic failures in current video generation architectures. To properly diagnose these failures, we categorized the evaluated baselines into two distinct groups: **Group 1: Foundation T2V Models** and **Group 2: Specialized T2MSV Frameworks**.

4.1 Group 1: Foundation T2V Models and the Static Trap

Foundation models (e.g., CogVideoX, LTX-Video, SVD) exhibit the raw limits of monolithic spatiotemporal priors. They achieve high Subject Consistency (> 0.80) but catastrophically low Background Diversity (< 0.10) in Track S. More importantly, their **DSA** scores remain near zero, proving they generate a single static scene regardless of narrative instructions. This results in a uniform probability distribution in the DSA heatmaps (Fig. 3(A)), indicating a total failure to isolate instructions.

4.2 Group 2: T2MSV Frameworks and Identity Amnesia

Specialized frameworks successfully break out of the Static Trap, achieving higher Background Diversity and DSA scores in Track S. However, this dynamism comes at the cost of identity preservation. Their Subject Consistency plummets, a phenomenon we define as **Identity Amnesia**. As visualized in our performance landscape in Fig. 5, these models prioritize dynamics but fail to maintain a coherent subject across shots.

4.3 The "Double-Kill" Pareto Frontier

Our comprehensive analysis uncovers a fundamental **"Double-Kill" dilemma**: current architectures either preserve identity at the expense of dynamics (the Static Trap) or prioritize dynamics but suffer from severe Identity Amnesia. As shown in the scatter plot in Fig. 5, models are clustered in either the top-left or bottom-right quadrants, leaving the top-right **"Unoccupied Ideal Zone"** completely empty. This Pareto frontier suggests a structural entanglement in current video priors that prevents simultaneous spatiotemporal control.

4.4 The Context Paradox

Our dual-track evaluation exposes the **Context Paradox**: a model's metric performance can be misleading without contextualizing the scenario. For instance, CogVideoX exhibits exceptionally low Background Diversity. In Track M, this might be misidentified as excellent spatial stability. However, the same model exhibits the same low diversity in Track S, proving it is merely trapped in static generation. This cross-scenario diagnosis is essential for distinguishing intentional continuity from generative failure.

5 Conclusion

In this paper, we have systematically exposed the fundamental flaws of current Text-to-Multi-Shot Video (T2MSV) evaluation methodologies. By introducing the **DIAL** (Diagonal Instruction ALignment) benchmark and its two orthogonal tracks—Track S (Semantic Leap) and Track M (Motion Continuity)—we have provided a rigorous diagnostic framework for identifying critical failure modes: the **Static Trap**, **Context Bleeding**, and **Identity Amnesia**.

Central to our work is the formulation of **Diagonal Semantic Alignment (DSA)**, which shifts the paradigm from absolute alignment to contrastive instruction isolation. By leveraging a mathematically robust, zero-sum formulation, DSA provides the first objective mechanism for penalizing instruction-agnostic generation and ensuring shot-level semantic independence. Our exhaustive evaluation of state-of-the-art models uncovers a pervasive **"Double-Kill" dilemma**, where current architectures struggle to balance identity preservation with narrative dynamism.

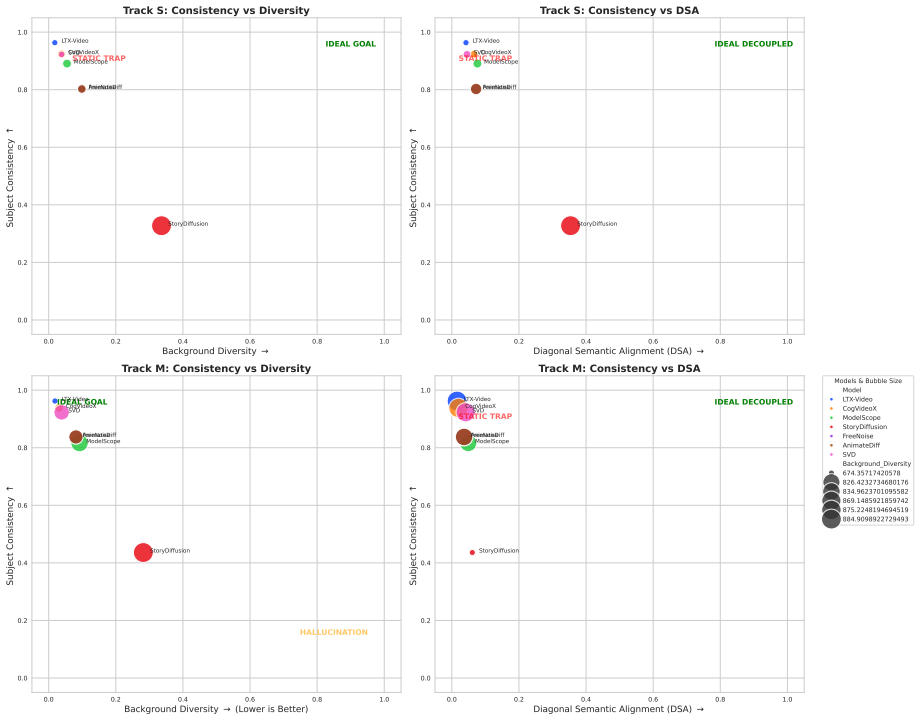


Figure 5: The comprehensive performance landscape across Track S and Track M. The plots explicitly show the relationship between Subject Consistency, Background Diversity, and our proposed Diagonal Semantic Alignment (DSA). Current models are completely missing from the "Ideal Decoupled" goal (high consistency, high DSA).

Ultimately, DIAL serves as more than a leaderboard; it is a critical diagnostic foundation for the next generation of video generative models. Our findings demonstrate that current monolithic architectures are structurally insufficient for mastering complex multi-shot narratives. The community must pivot toward **"Decoupled Dynamics"**—architectures that explicitly isolate identity-preserving priors from narrative-driven motion priors. We hope this benchmark will catalyze research into more robust, spatiotemporally decoupled video priors, paving the way for truly cinematic, instruction-aligned video synthesis.