VISTA: A Test-Time Self-Improving Video Generation Agent

Do Xuan Long[1 2 *], Xingchen Wan[1], Hootan Nakhost[1], Chen-Yu Lee[1], Tomas Pfister[1] and Sercan Ö. Arık[1]

[1]Google, [2]National University of Singapore

Despite rapid advances in text-to-video synthesis, generated video quality remains critically dependent on precise user prompts. Existing test-time optimization methods, successful in other domains, struggle with the multi-faceted nature of video. In this work, we introduce VISTA, a novel multi-agent system that autonomously improves video generation through refining prompts in an iterative loop. VISTA first decomposes a user's idea into a structured temporal plan. After generation, the best video is identified through a robust pairwise tournament. This winning video is then critiqued by a trio of specialized agents focusing on visual, audio, and contextual fidelity. Finally, a reasoning agent synthesizes this feedback to introspectively rewrite and enhance the prompt for the next generation cycle. Experiments on single- and multi-scene video generation scenarios show that while prior methods yield inconsistent gains, VISTA consistently improves video quality and alignment with user intent, achieving up to 60% pairwise win rate against state-of-the-art baselines. Human evaluators concur, preferring VISTA's outputs in 66.4% of comparisons.

Direct Prompting (DP)	**VISTA (Ours)**

Single-scene (Polyak et al., 2025): The person's forehead creased with worry as he listened to bad news.

Single-scene (Polyak et al., 2025): A spaceship entering hyperdrive, stars streaking past as it accelerates.

Multi-scene-Interviews: The video features a man outdoors, asking a trivia question about a comedian known...

Multi-scene-Animation: The video is an educational animation designed for young children...

Table 1 | Example videos generated by VISTA, showing improvements in visual fidelity, camera focus, smooth transitions, compelling storylines, and sounds. Optimized prompts and more examples are in our project page.

Project page: https://g-vista.github.io/

1. Introduction

Text-to-video (T2V) generation has seen significant progress, with models like Veo 3 (Google Deepmind, 2025) demonstrating a remarkable ability to generate coherent, high-quality video and audio

---

Corresponding author(s): xuanlong.do@u.nus.edu, xingchenw@google.com, soarik@google.com
* This work was done while Do Xuan Long was a Student Researcher at Google.

from text. This progress positions T2V as a powerful tool for applications in creative storytelling, education, and content creation (Liu et al., 2024b). However, broader deployment is limited by several persistent challenges. Models often struggle with precise alignment with user goals, consistent adherence to physical laws and commonsense (Bansal et al., 2024), and a high sensitivity to the exact phrasing of input prompts. This sensitivity, in particular, forces users into a laborious trial-and-error cycle, requiring them to repeatedly tweak phrasing and filter outputs to achieve a desirable result.

Meanwhile, test-time optimization frameworks have shown promise in automatically improving generation quality and preference alignment for both text and image generation (Hao et al., 2023; Madaan et al., 2023; Muennighoff et al., 2025; Pryzant et al., 2023; Snell et al., 2025). However, extending them to autonomous video generation presents substantially challenges. Unlike text or image data, videos unfold across multiple scenes, modalities, and high-level contextual meaning, making evaluation and optimization significantly more complex. Existing efforts often target only specific video properties, such as objects (Gao et al., 2025a), harmless-accurate-helpful traits (Cheng et al., 2025b), or visual-reward fine-tuning (Dalal et al., 2025; Ji et al., 2024; Soni et al., 2024). Yet, to the best of our knowledge, no studies have unified visual, audio, and contextual quality in a single optimization framework, despite these together being key to user satisfaction.

We introduce VISTA (Video Iterative Self-improvemenT Agent), a novel multi-agent framework that self-improves video-audio generation at test time. Inspired by how humans evaluate videos and refine prompts, VISTA jointly optimizes three key aspects of videos: Visual, Audio, and Context, through collaborative agents. This process is guided by a comprehensive and configurable suite of evaluation metrics tailored to each, and is driven by four key components: **(i) Structured Video Prompt Planning** (Section 2.1), which transforms user input into temporally grounded, multi-scene, and multi-aspect descriptions; **(ii) Pairwise Tournament Selection** (Section 2.1), a probing-driven algorithm to identify the best video-prompt candidates; **(iii) Multi-Dimensional Multi-Agent Critiques** (Section 2.2), a triadic agent system that provides nuanced critiques motivated by the Jury Decision Process (Klevorick and Rothschild, 1979); and a **(iv) Deep Thinking Prompting Agent**, which performs human-like introspective, structured reasoning to revise prompts in a targeted way. We rigorously evaluate VISTA on a widely used single-scene benchmark and an internal multi-scene benchmark with rich instructions. Experiments show that VISTA substantially outperforms prior test-time optimization methods, improving state-of-the-art (SOTA) T2V models like Veo 3 by up to 60% under our metrics, further validated by human evaluations. Our work enables more reliable, user-aligned text-to-video generation and paves the way for broader video synthesis applications. In summary, this paper makes the following contributions:

- We propose VISTA, a novel multi-agent framework that emulates human-like prompt refinement to improve T2V generation. To the best of our knowledge, VISTA is the first to jointly improve the visual, audio, and context dimensions of videos. See Table 1 for its exemplary generated videos.

- We develop VISTA's components and meticulously design their configurable evaluation metrics that enable fully autonomous, model-driven video evaluation and refinement.

- We conduct extensive experiments supported by in-depth analysis and human studies showing that VISTA consistently outperforms existing baselines and improves user preferences.

2. VISTA

VISTA (Figure 1 and Alg. 1) is a modular, configurable framework for optimizing text-to-video generation. Given a user video prompt P, it produces an optimized video V^* and its refined prompt P^* through two phases: **(i) Initialization** and **(ii) Self-Improvement**, inspired by the human video
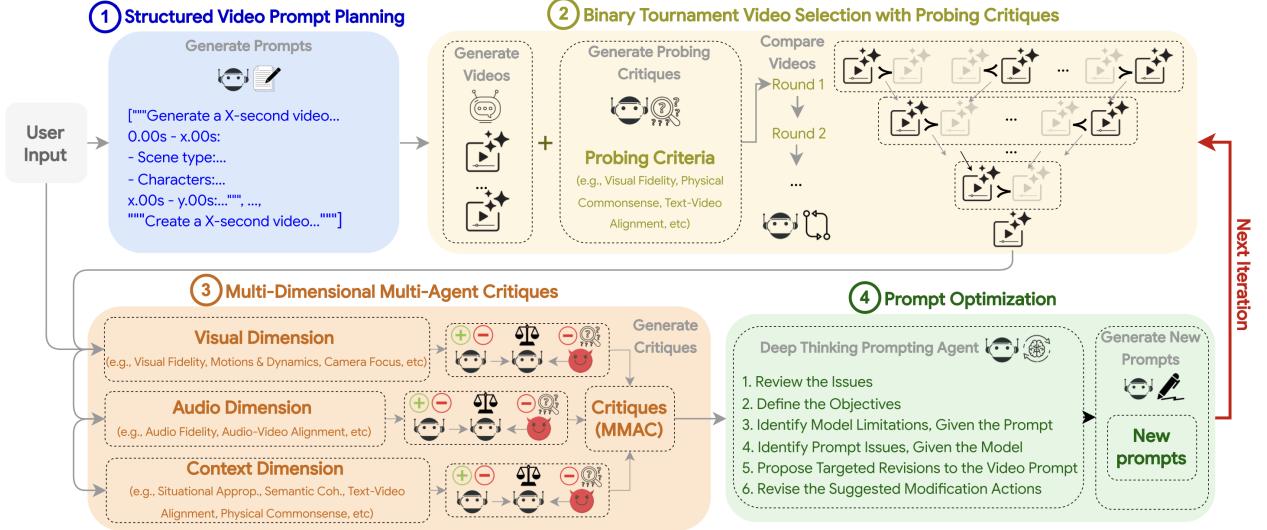
Figure 1 | The workflow of our proposed multi-agent framework, VISTA. 📷: MLLM Agent; 😈: Adversarial MLLM Agent; 🤖: Video Generation Agent.

optimization process via prompting. During (i), the prompt is parsed and planned into variants to generate candidate videos (**Step 1**), after which the best video-prompt pair is selected (**Step 2**). In (ii), the system generates multi-dimensional, multi-agent critiques (**Step 3**), refines the prompt (**Step 4**), produces new videos, and reselects the champion pair (**Step 2**). This phase continues until a stopping criterion is met or the maximum number of iterations is reached.

2.1. Initialization Phase

Step 1. Structured Video Prompt Planning (Alg. 1-L1). The user video generation prompt P is first parsed into m **timed sequences of scenes** as prompt candidates where each candidate is defined as $P_i := [S_{i,1}, S_{i,2}, \dots]$. By default, each scene configuration consists of nine properties spanning context, visual, and audio dimensions: **(1) Duration:** The length (in seconds) of the scene; **Scene Type:** The type of the scene (e.g., action, montage); **(3) Characters:** Any key figures or entities on screen, whether human, animal, or object, that drive the scene (e.g., a dog, a flower); **(4) Actions:** Any deliberate and meaningful movements or behaviors that the characters above perform (e.g., the flower is blooming); **(5) Dialogues:** Any spoken lines, voiceovers, or on-screen text that drive the scene; **(6) Visual Environment:** Any environmental backdrop or setting that establishes the scene's world (e.g., a tranquil canopy of boundless sky); **(7) Camera:** Any cinematographic technique, including framing, movement, or angle, that shapes the scene (e.g., a close-up capturing subtle emotion); **(8) Sounds:** Any auditory elements, such as soundtrack, voiceovers, or ambient noises, that enrich the scene (e.g., a crashing waves in the background); **(9) Moods:** Any prevailing tone or atmosphere that powers the scene (e.g., serene).

VISTA implements a configurable interface over the above properties and planning constraints on scenes (see below). This introduces two main novelties over prior work (Google Cloud, 2024; Huang et al., 2024; Polyak et al., 2025): **(i) temporal scene-level decomposition**, where user prompts are structured as semantically coherent segments to support reasoning over complex content, and **(ii) fine-grained multi-modal video prompting**, enabling automated multi-dimensional critiques and self-improvement. A multi-modal LLM (MLLM) is used to obtain $S_{i,j}$ and, when needed, infer missing scene properties. To preserve fidelity to the user's intent and avoid scenario drift, the system by default enforces planning constraints on *Realism, Relevancy, and Creativity*, though they are configurable

---

Algorithm 1 VISTA: Configurable Self-Improving Video Generation Agent

---

Require: User prompt P, a text-to-video model T2V(), a multi-modal MLLM, #iteration T, optional early-stop m, optional configurable criteria for video selection $\mathcal{M}_{user}^S$ and critiques $\mathcal{M}_{user}^C$
Ensure: Optimized video output V^* with its prompt P^*

    [Initialization Phase]
  1: Plan and generate video prompts $\mathcal{P} := \{P_1, \ldots, P_m, P\} \leftarrow \textsc{PromptPlanner}(P)$
  2: Generate video candidates $\mathcal{V} := \{V_1, \ldots, V_n\}$, where $\{V_i, \ldots\} \leftarrow \text{T2V}(P_j)$
  3: Select best video-prompt pair $(V^*, P^*) \leftarrow \textsc{PairwiseSelect}(\mathcal{V}, \mathcal{P}, P, \mathcal{M}_{user}^S)$
    [Self-Improvement Phase]
  4: **for** $t = 1$ to T **do**
  5:    Generate multi-dimensional multi-agent critiques $\mathcal{F}^t \leftarrow \text{MMAC}(V^*, P^*, P, \mathcal{M}_{user}^C)$
  6:    Optimize prompts $\mathcal{P}^t := \{P_1^t, \ldots, P_m^t, P^*\} \leftarrow \textsc{PromptOptimizer}(P^*, P, \mathcal{F}^t)$
  7:    Generate video candidates $\mathcal{V}^t := \{V_1^t, \ldots, V_n^t, V^*\}$, where $\{V_i^t, \ldots\} \leftarrow \text{T2V}(P_j^t)$
  8:    Update best video-prompt pair $(V^*, P^*) \leftarrow \textsc{PairwiseSelect}(\mathcal{V}^t, \mathcal{P}^t, P, \mathcal{M}_{user}^S)$
  9:    **if** (V^*, P^*) does not change after m iterations **then**
10:        **break**
11:    **end if**
12: **end for**
13: **return** (V^*, P^*)

---

and optional: videos are grounded in real-world physics unless the prompt specifies otherwise (e.g., animated or fantastical); only elements explicitly stated or implied are included, avoiding unnecessary invention; ambient sounds and effects are encouraged when beneficial; and excessive scene transitions are discouraged for simple or short prompts. See Section B.1.1 for defaults. Finally, we retain a residual P in the sampled set $\mathcal{P}$ to allow models that do not benefit from decomposition.

Step 2. Binary Tournament Video Selection with Probing Critiques (Alg. 1-L3, L10). After obtaining candidate video-prompt pairs, this step selects the pair (V^*, P^*) with the highest video quality for self-improvement. Conventional video evaluation methods often rely on complex metric-based systems (Huang et al., 2024), which are computationally expensive in practice (Zhang et al., 2025). Inspired by recent advances demonstrating the strong video understanding capabilities of MLLMs (Fu et al., 2025; Li et al., 2024b), we employ an MLLM-as-a-Judge to evaluate videos on customizable evaluation criteria $\mathcal{M}_{user}^S$. However, scoring videos without ground-truth references is inherently subjective and unreliable. We adopt a video pairwise comparison strategy, which better aligns with human preferences in reinforcement learning and language tasks, while avoiding model-induced biases (Lee et al., 2024; Liu et al., 2024c). As detailed in Alg. 2, we iteratively reduce the set of candidate videos via **Binary Tournaments** (Miller et al., 1995). At each iteration, videos are grouped into pairs and compared bidirectionally via swapping; only the winning videos advance to the next round. We swap the videos to avoid token biases (Zheng et al., 2024a).

Comparing (V_i, V_j), however, remains unreliable to align with human preferences. To improve this, we employ a criteria-based approach (Li et al., 2024a) where the model judges videos across metrics $\mathcal{M}_{user}^S$. We meticulously design a default configuration where $\mathcal{M}_{user}^S := \{$**Visual Fidelity; Physical Commonsense; Text-Video Alignment; Audio-Video Alignment; Engagement**$\}$ capturing the core aspects of video quality, and the video winning on more criteria is the winner. While this approach is more principled, we still observe that model often fails to provide sufficiently critical assessments. We attribute this to the dual burden placed on the model: analyzing videos and comparing them simultaneously. To mitigate this, we introduce a two-step decomposition: first, for each video, the

---

Algorithm 2 Pairwise Tournament Selection (PAIRWISESELECT)

Require: Video prompt P, list of videos $\mathcal{V} = \{V_1, \ldots, V_n\}$ and their prompts $\mathcal{P} = \{P_1, \ldots, P_n\}$, and optional configurable criteria $\mathcal{M}_{user}^S$

Ensure: Best video V^* and its prompt P^*

 1: Generate probing critiques $\mathcal{Q} := \{Q_1, \ldots, Q_n\} \leftarrow \text{MLLM}(\mathcal{V}, \mathcal{P}, \mathcal{M}_{user}^S)$
 2: **while** $|\mathcal{V}| > 1$ **do**
 3:      Group $\mathcal{V}$ into pairs.
 4:      **for** any pair (V_i, V_j) **do**
 5:          $\left(V_{win}^f, V_{lose}^f\right) \leftarrow \text{MLLM}(V_i, Q_i, V_j, Q_j, \mathcal{M}_{user}^S)$ ▷ Forward pairwise comparison.
 6:          $\left(V_{win}^s, V_{lose}^s\right) \leftarrow \text{MLLM}(V_j, Q_j, V_i, Q_i, \mathcal{M}_{user}^S)$ ▷ Swapped pairwise comparison.
 7:          $(V_{win}, V_{lose}) \leftarrow \left(V_{win}^f, V_{lose}^f\right)$ if $\left(V_{win}^f, V_{lose}^f\right) == (V_{win}^s, V_{lose}^s)$ else assign randomly
 8:          $\mathcal{V}.\text{remove}(V_{lose})$; $\mathcal{P}.\text{remove}(P_{lose})$
 9:      **end for**
10: **end while**
11: **return** $(\mathcal{V}[0], \mathcal{P}[0])$.

---

model generates probing critiques ($\mathcal{Q}$ in Alg. 2-L1) on $\mathcal{M}_{user}^S$, then these critiques are used to support the comparisons. Finally, we implement customizable penalty mechanisms for T2V failures during selection: by default, VISTA penalizes common failures in $\mathcal{M}_{user}^S$ unless the user prompt explicitly specifies otherwise. The final scores of video candidates V_i and V_j are s_i and s_j:

$$s_i \leftarrow \frac{1}{k} \Sigma_{C \in \mathcal{M}_{user}^S} \left(\delta(C, V_i, V_j) - \lambda \cdot \mathbb{1}(C, V_i)\right), \quad s_j \leftarrow \frac{1}{k} \Sigma_{C \in \mathcal{M}_{user}^S} \left(1 - \delta(C, V_i, V_j) - \lambda \cdot \mathbb{1}(C, V_j)\right)$$

where $\delta(C, V_i, V_j) \in \{0, 0.5, 1\}$ represents the outcome of V_i against V_j on C, corresponding to {Loss, Tie, Win}, and $\mathbb{1}(C, V) \in \{0, 1\}$ indicates whether V violates C to some extent. The term λ is a penalty applied for violations. Noteworthily, $\mathbb{1}(C, V)$ can be customized to any preferred constraints, not necessarily restricted to $\mathcal{M}_{user}^C$. See Section B.2 for our default prompts and metrics definitions.

2.2. Self-Improvement Phase

Step 3. Multi-Dimensional Multi-Agent Critiques (Alg. 1-L5). Given (P, V^*, P^*), this step elicits targeted critiques to refine the prompt P^*. Video evaluation is difficult due to its multi-dimensional nature. We address this with a multi-agent critique framework decomposed into $\mathcal{D} = \{\textbf{Visual, Audio, Context}\}$, each assessed independently by a dedicated system. Evaluation criteria are configured through $\mathcal{M}_{user}^C$. While prior work (Liu et al., 2024a; Zheng et al., 2024b) proposed diverse metrics, recent SOTA models (Google Deepmind, 2025; Wan et al., 2025a) already excel on most of them, suggesting they insufficiently differentiate between high-quality generations (see Section 4.1). Therefore, we carefully design a comprehensive default configuration for $\mathcal{M}_{user}^C$ that expose modality-specific failures even in SOTA T2V models. These are strategically selected and refined from (Bansal et al., 2024; Cheng et al., 2025a; Gao et al., 2023; Liu et al., 2024a):

- **Visual**: Visual Fidelity, Motions and Dynamics, Temporal Consistency, Camera Focus, Visual Safety.

- **Audio**: Audio Fidelity, Audio-Video Alignment, Audio Safety.

- **Context**: Situational Appropriateness, Semantic Coherence, Text-Video Alignment, Physical Commonsense, Engagement, Video Format (Beginning, Ending, Transitions).

These include both human-centric criteria such as situational appropriateness and video format, and fine-grained video metrics like visual fidelity[1]. However, we observe that directly employing MLLM-as-a-Judge often yields shallow and unuseful critiques (even when being explicitly asked to be critical). This is because SOTA T2V models such as Veo 3 already produce high-quality outputs that are difficult to critique at surface level even by humans. To address this gap, we introduce **Multi-Dimensional Multi-Agent Critiques (MMAC)**: inspired by the **Jury Decision Process** (Klevorick and Rothschild, 1979), for each evaluation dimension $D \in \mathcal{D}$, we construct a triadic court consisting of a **Normal Judge** that critically assesses and scores the video on D's metrics in both good and bad faiths, an **Adversarial Judge** which generates probing questions, counterarguments, and scores to expose video flaws on D's metrics, and a **Meta Judge** that consolidates the judges from both sides:

$$
\begin{aligned}
\{C_D, S_D\} &\leftarrow J_D\left(P, V^*, P^*\right) \quad \text{(Normal Judge)} \\
\{C_D^-, S_D^-\} &\leftarrow J_D^-\left(P, V^*, P^*\right) \quad \text{(Adversarial Judge)} \\
\{C_D^*, S_D^*\} &\leftarrow J_D^*\left(P, C_D, S_D, C_D^-, S_D^-\right) \quad \text{(Meta Judge)}
\end{aligned}
\tag{1}
$$

The MMAC is $\mathcal{F} := \{C_D^*, S_D^* | D \in \mathcal{D}\}$ where C_D and S_D are metrics' critiques and scores (on a scale of 1-10 (Zheng et al., 2023)). See Section B.3 for default prompts with metrics' definitions.

Step 4. Prompt Optimization (Alg. 1-L6). After obtaining $\mathcal{F}$, this step refines P^* via a **Deep Thinking Prompting Agent (DTPA)**. Direct MLLM optimization often overcomplicates prompts and interprets critiques shallowly (see Section 4.3 for examples); DTPA instead performs a six-step, self-reflective reasoning to suggest prompt modifications in one chain-of-thought: (1) identifying video issues via metrics with low meta scores (≤ 8), (2) clarifying the expected outcome and success criteria, (3) evaluating the context sufficiency in the current prompt, (4) determining whether failures stem from model limitations or prompt, and (5) detecting potential conflicts or vagueness within the prompt itself. Based on this introspective analysis, the DTPA proposes a set of modification actions, which are then (6) reviewed and refined to ensure they fully address the failures identified in (1). See Sections D.1 and D.5 for VISTA's optimized prompts and Section B.5 for method's prompts.

$$
\mathcal{M} := \{M_1, \dots\} \leftarrow \text{DTPA}(P, P^*, \mathcal{F}) \quad (M_i \text{ are suggested modifications})
\tag{2}
$$

These modifications are then used to sample improved prompts:

$$
\mathcal{P} := \{P_1, \dots, P_n, P^*\} \leftarrow \text{MLLM}(P, P^*, \mathcal{M})
\tag{3}
$$

3. Related Work

3.1. Text-To-Video Synthesis and Optimization

Recent years have seen major advances in T2V synthesis (Google Deepmind, 2025; Hong et al., 2023; OpenAI, 2024; Polyak et al., 2025; Wan et al., 2025a), with SOTA models such as Sora and Veo 3 receiving widespread attention–Veo 3 notably pioneering high-quality audio-video generation. Yet, current models remain highly prompt-sensitive. Existing prompt optimization and test-time self-evolving (agentic) methods, while being successful in other domains (Fang et al., 2025; Gao et al., 2025b; Long et al., 2025; Mañas et al., 2024; Schulhoff et al., 2024; Wan et al., 2025b), are limited in video optimization, often requiring white-box access or fine-tuning. For example, VideoAgent (Soni

---

[1]This set of metrics is more comprehensive than in Step 2 because this step prioritizes in-depth and granular critiques.

et al., 2024) refines video plans by executing them online and using successful trajectories to fine-tune the generation model, while MotionPrompt (Nam et al., 2025) learns token embeddings to improve motion fidelity, and RAPO (Gao et al., 2025a) rewrites and augments prompts but relies on target prompts during training. Closest to our work, VPO (Cheng et al., 2025b) optimizes for harmlessness, accuracy, and helpfulness, but not at test time. Meanwhile, LM-powered multi-agent systems such as Mora (Yuan et al., 2024) and FilmAgent (Xu et al., 2025) tackle tasks such as scriptwriting and cinematography, yet omit test-time optimization. VISTA combines these directions and is the first to explore black-box prompt optimization for video generation.

3.2. Video and Audio Generation Evaluation

Video generation is uniquely challenging to evaluate: there is no single, definitive "ground truth", and videos inherently span multiple dimensions that require complex and holistic reasoning to evaluate effectively. Conventional visual quality metrics, such as Inception Score (IS) (Salimans et al., 2016), FID (Heusel et al., 2017), and CLIP-Score (Hessel et al., 2021), typically operate along a single dimension and do not to provide a comprehensive evaluation. Recent video benchmarks such as T2I-CompBench (Huang et al., 2023), VBench (Huang et al., 2024), and EvalCrafter (Liu et al., 2024a) offer multi-dimensional evaluations tailored to specific model capabilities. However, they still heavily rely on conventional single-metric measures that are time-intensive, inefficient for autonomous improvement, and exclude audio evaluation. Audio-visual extensions like TAVGBench (Mao et al., 2024) and ACVUBench (Yang et al., 2025) illustrate complementary but rigid approaches: the former emphasizes embedding similarity without failure-focused reasoning, while the latter focuses on evaluating MLLM understanding capabilities. Closest to our multi-agent critiques approach are VideoScore (He et al., 2024) and Evaluation Agent (Zhang et al., 2025), nevertheless, they are not failure-focused and overlook audio dimensions. VISTA emphasizes failure-sensitive visual and audio metrics even for SOTA T2V models, aiming to efficiently enhance AI-generated videos.

4. Experiments

Benchmarks. We evaluate our approach and baselines on two benchmarks representing distinct scenarios: **single-scene** and **multi-scene** generation. For single-scene evaluation, we follow Dalal et al. (2025) to use MovieGenVideo (Polyak et al., 2025) benchmark via randomly selecting 100 prompts. For multi-scene evaluation, we use 161 prompts with at least two scenes from our internal dataset covering diverse topics and matching the T2V model's duration requirements.

Models and Baselines. VISTA employs two core components: a multimodal large language model (MLLM) and a text-to-video generation model (T2V). For our experiments, we use **Gemini 2.5 Flash** (`Gemini-2.5-flash-preview-05-20`) (Gemini Team, 2025) as the MLLM and **Veo 3** (`Veo-3.0-generate-preview`) (Google Deepmind, 2025) as the video generator because they are among the current state of the art. We also experiment with a less powerful T2V model (**Veo 2** (Google Deepmind, 2024)) in Section 4.3. Since there is almost no test-time prompt optimization or multi-agent system for video generation up-to-date, we compare VISTA with four baselines: **(1) Direct Prompting (DP)**, which directly uses the user prompt as input for video generation; **(2) Visual Self-Refine (VSR)** (Madaan et al., 2023)[2], which leverages the MLLM to iteratively evaluate the video generated by the T2V model and subsequently refine the prompt; **(3) Rewrite (Google Cloud, 2024)**, which involves using the MLLM to rewrite the user prompt using the Vertex AI video

---

[2]https://github.com/madaan/self-refine/blob/main/colabs/Visual-Self-Refine-GPT4V.ipynb

	Init				2				3				4				5			
Method	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ
Baseline vs. Direct Prompting																				
VSR	22.4	61.2	16.4	6.0	34.9	42.4	22.7	12.2	23.4	48.4	28.1	-4.7	26.6	51.6	21.9	4.7	24.6	47.7	27.7	-3.1
VSR++	28.6†	67.3†	4.1†	24.5†	31.1†	42.2†	26.7†	4.4†	29.5†	52.3†	18.2†	11.3†	26.8†	46.3†	26.8†	0.0†	33.3†	53.3†	13.3†	20.0†
Rewrite	19.0	69.0	12.0	7.0	35.0†	52.0†	13.0†	22.0†	30.0†	56.0†	14.0†	16.0†	39.0†	47.0†	14.0†	25.0†	27.0†	65.0†	8.0†	19.0†
VPO	29.0	46.0	25.0	4.0	34.0†	56.0†	10.0†	24.0†	31.0†	55.0†	14.0†	17.0†	28.0†	65.0†	7.0†	21.0†	36.0†	56.0†	8.0†	28.0†
VISTA	**35.5**	50.1	14.4	**21.1**	**40.7**	49.4	9.9	**30.8**	**41.4**	43.7	14.9	**26.5**	**42.4**	43.5	14.1	**28.3**	**45.9**	50.2	13.9	**32.0**
VISTA vs. Baselines																				
VSR	47.0	36.4	16.7	30.3	60.0	35.0	5.0	55.0	45.6	43.9	10.5	35.1	38.6	47.4	14.0	24.6	45.6	35.1	19.3	26.3
VSR++	30.4†	51.1†	18.5†	11.9†	48.8†	36.2†	15.0†	33.8†	42.3†	42.3†	15.4†	26.9†	50.7†	29.6†	19.7†	31.0†	30.3†	57.9†	11.8†	18.5†
Rewrite	34.0	51.6	14.4	19.6	34.8†	45.7†	19.6†	15.2†	35.6†	54.4†	10.0†	25.6†	38.2†	42.7†	19.1†	19.1†	40.2†	41.4†	18.4†	21.8†
VPO	27.8	66.7	5.6	22.2	43.8†	50.0†	6.2†	37.6†	40.0†	60.0†	0.0†	40.0†	53.3†	40.0†	6.7†	46.6†	35.7†	45.7†	18.6†	17.1†

(a) Results in single-scene scenarios.

	Init				2				3				4				5			
Method	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ	Win	Tie	Loss	Δ
Baseline vs. Direct Prompting																				
VSR	21.2	63.5	15.4	5.8	33.3	40.4	26.3	7.0	36.7	49.0	14.3	22.4	33.7	52.0	14.3	19.4	35.3	45.9	18.8	16.5
VSR++	29.4†	41.2†	29.4†	0.0†	26.5†	55.9†	17.6†	8.9†	17.6†	58.8†	23.5†	-5.9†	23.5†	52.9†	23.5†	0.0†	26.5†	52.9†	20.6†	5.9†
Rewrite	14.3	55.1	30.6	-16.3	31.8†	55.7†	12.5†	19.3†	29.6†	48.9†	21.6†	8.0†	23.9†	55.7†	20.5†	3.4†	23.9†	52.3†	23.9†	0.0†
VPO	25.2	53.7	21.1	4.1	38.2†	52.8†	9.0†	29.2†	38.2†	51.7†	10.1†	28.1†	28.1†	60.7†	11.2†	16.9†	27.0†	60.7†	12.4†	14.6†
VISTA	**37.8**	52.3	9.9	**27.9**	**39.4**	47.2	13.4	**26.0**	**38.4**	52.2	9.4	**29.0**	**43.7**	43.7	12.6	**31.1**	**46.3**	42.5	11.2	**35.1**
VISTA vs. Baselines																				
VSR	46.2	44.0	9.9	36.3	43.8	43.8	12.5	31.3	53.2	36.7	10.1	43.1	50.0	40.3	9.7	40.3	48.5	37.9	13.6	34.9
VSR++	35.3†	52.9†	11.8†	23.5†	43.8†	46.9†	9.4†	34.4†	35.0†	46.9†	18.1†	16.9†	34.4†	53.1†	12.5†	21.9†	34.4†	50.0†	15.6†	18.8†
Rewrite	33.5	59.5	8.0	25.5	32.5†	67.5†	0.0†	32.5†	36.6†	57.3†	6.1†	30.5†	37.0†	58.0†	4.9†	32.1†	42.0†	51.9†	6.2†	35.8†
VPO	25.3	64.3	11.4	13.9	27.7†	69.9†	2.4†	25.3†	34.2†	48.8†	17.1†	17.1†	18.5†	74.1†	7.4†	11.1†	25.9†	60.5†	13.6†	12.3†

(b) Results in multi-scene scenarios.

Table 2 | Win/Tie/Loss rates and Δ = Win − Loss across 5 iterations. † refers to our scaled-up results, and underlines are results evaluated on half of the benchmark.

generation prompt guidelines provided by Google; **(4) VPO (Cheng et al., 2025b)**, which expands the user prompts based on three core principles of harmlessness, accuracy, and helpfulness.

We run VISTA for 5 iterations: 1 initialization followed by 4 self-improving steps. In both phases, we sample 5 prompts, each with 3 variants, and generate 2 videos per prompt, resulting in 30 videos per iteration. Since Rewrite and VPO operate in a single iteration, we scale their #videos by matching these used in VISTA's four self-improvement iterations. Likewise, we scale VSR to match VISTA's total #videos, denoted as **Visual Self-Refine++ (VSR++)**. For scaled baselines, the best video per iteration is chosen via binary tournament with bidirectional pairwise comparisons (Section B.7).

4.1. Automatic Evaluations

Main Evaluations. We evaluate using both MLLM-as-a-Judge and conventional metrics. Following Dalal et al. (2025), we conduct pairwise comparisons between our method and baselines, using Gemini 2.5 Flash (Gemini Team, 2025) as the evaluator for its SOTA video-audio understanding and multi-video processing. We assess ten criteria: **Visual Fidelity, Motions, Temporal Consistency (scene level), Text-Video Alignment, Audio Quality, Audio-Video Alignment, Situational Appropriateness, Physical Commonsense, Engagement, and Video Format (beginning, ending, transitions)**. Each comparison is bidirectional, with outcomes recorded as **Win/Tie/Loss** and Δ = **Win-Loss**; conflicts after swapping are marked as Ties. A video wins if it excels in at least three criteria and does not lose on Text-Video Alignment. See Section B.6 for prompts and metric definitions. We also report conventional metrics: **IS** (Salimans et al., 2016), **CLIP-Score** (Hessel et al., 2021), eight visual metrics from **VBench** (Huang et al., 2024), and three audio metrics from **NISQA** (Mittag et al., 2021). Additional results with other evaluators are in Sections A.6 and A.7.
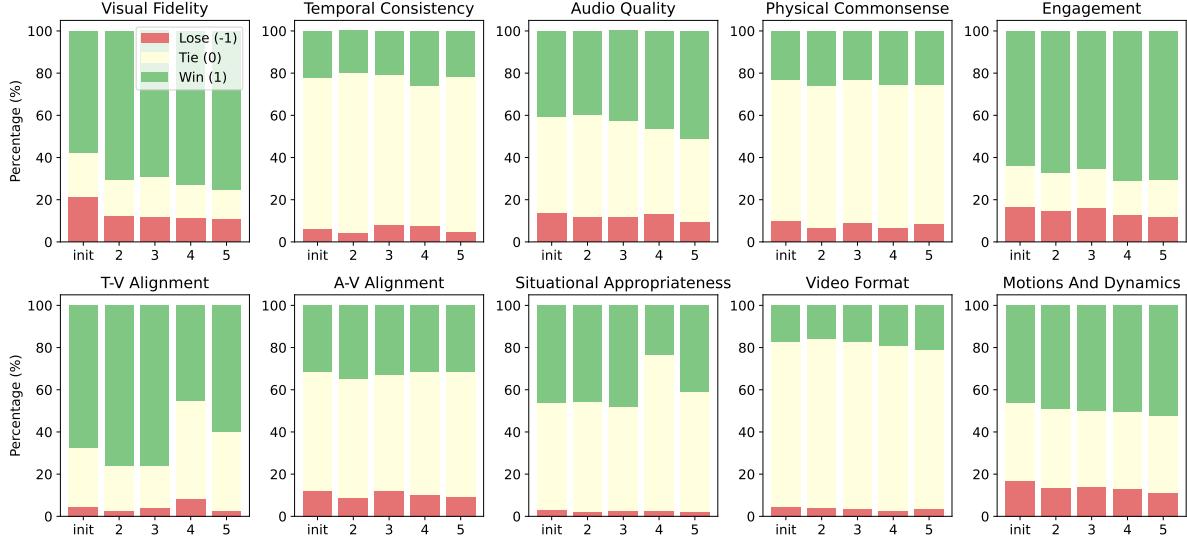
Figure 2 | Average win/tie/lose between VISTA and Direct Prompting (DP) on two benchmarks. The individual benchmark results are in Section A.5

Baselines vs. DP. Results. From Table 2, the original baselines (without †) show inconsistent and often unfavorable win-loss dynamics. For example, in Table 2-(a), VPO reaches 29% win rate with only $\Delta = 4.0\%$, while in the multi-scene setting (Table 2-(b)), Rewrite even yields $\Delta = -16.3\%$. This highlights their limited robustness, where gains in some aspects trade off against regressions elsewhere. VSR performs better in multi-scene but adds little in single-scene scenarios, likely because the former is more challenging: DP underperforms in multi-scene, giving VSR room to help, whereas DP's stronger single-scene performance restricts its impact. Enhanced baselines (with †) mitigate this somewhat by sampling multiple videos and selecting the best, but show no evidence of continued improvement when scaling # videos more. By contrast, VISTA consistently outperforms all baselines and scales more effectively with test-time compute, achieving substantial gains with win rates up to 45.9% ($\Delta = 32\%$) and 46.3% ($\Delta = 35.1\%$).

VISTA vs. Baselines Results. VISTA achieves significant win rates over baselines, ranging from 27.8-60.0% (single-scene) and 18.5-53.2% (multi-scene). Notably, these trends do not correlate with baseline performance relative to DP. For example, while VSR performs best against DP, VISTA surpasses it by the largest margin. This likely reflects the VISTA's Pareto optimization of multiple video dimensions, whereas baselines typically focus on fewer aspects with narrower coverage.

Fine-Grained Results. To better understand the specific metrics where VISTA optimizes, we present a fine-grained pairwise comparison between VISTA and DP in Figure 2. Results reveal that VISTA achieves significant and consistent improvements across key visual, audio, and context dimensions. The most notable gains appear in Visual Fidelity, Engagement, Text-Video Alignment, Motions and Dynamics, Audio Quality, and Situational Appropriateness, attributable to VISTA's multi-dimensional critique framework and stringent selection process with constraints and penalties. Finally, VISTA yields only marginal improvements in Temporal Consistency and Video Format, as the Veo 3 model already demonstrates strong performance in these aspects.

Conventional Metrics Results. The results on conventional metrics are plotted in Figure 3. While prior frameworks exhibit no clear improvements on both visual and audio criteria over DP, we find
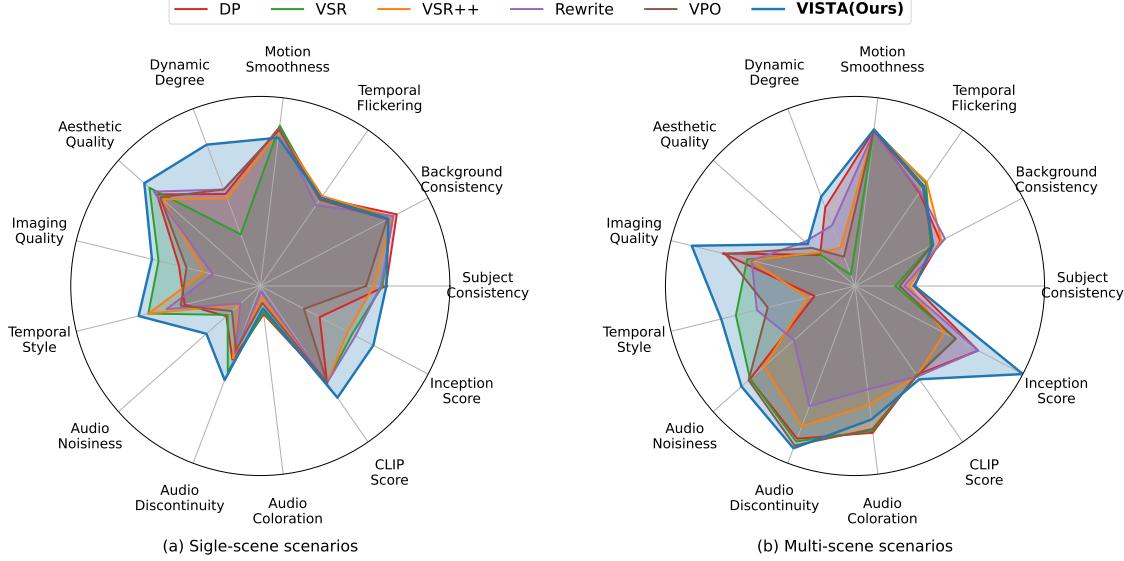
Figure 3 | Evaluation results using conventional metrics on single-scene (a) and multi-scene (b) benchmarks. Numerical results are provided in Tables 5 and 6.

that VISTA delivers notable improvements across key visual dimensions, particularly in Dynamic Degree, Aesthetic Quality, and Temporal Style. For example, it achieves 89.87% in Dynamic Quality in single-scene scenarios, surpassing the best-performing baseline at 77.22%, highlighting the VISTA's effectiveness in optimizing videos with richer and more realistic motion dynamics. In addition, it also reduces audio noisiness (+0.1/5 abs.) and discontinuities (+0.11/5 abs.) across both datasets, and delivers significant gains on CLIP-Score (+3% abs.) versus the best baselines. These gains closely align with the MLLM-as-a-Judge evaluations: both highlight VISTA's balanced improvements across visual fidelity, audio quality, and alignment.

4.2. Human Evaluations

Main Evaluations. We conduct three human evaluations to assess: (1) if humans prefer VISTA's outputs over the strongest baselines, (2) to what extent VISTA improves video quality across iterations, and (3) to what extend VISTA improves DP on visual and audio quality. For (1), we randomly sample 50 prompts, 25 from single-scene (VISTA vs. VSR++ at iteration 5) and 25 from multi-scene scenarios (VISTA vs. VSR at iteration 5) and employ five annotators with prompt optimization experience to label each pair as a Win or Loss. For (2), three different expert annotators score those 50 full optimization trajectories of both VISTA and VSR, using a 1-5 scale. For (3), we use the same 50 prompts and ask three expert annotators to rate videos on two specific dimensions: Visual Quality and Audio Quality using a 1-5 scale. See Section C.1 for our instructions.

Results. From the human evaluation results shown in Figure 4, VISTA consistently outperforms baselines across all metrics. The win rate comparison in (b) demonstrates its superiority with 66.4% versus 33.6% for best baseline. The self-improvement evaluation in (a) validates that VISTA indeed achieves meaningful self-improvement with an average score
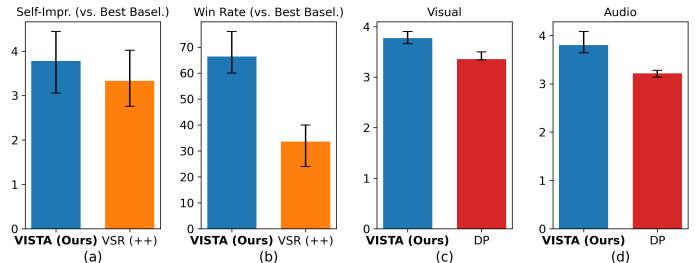

Figure 4 | Summary of human results. The individual annotators' results are in Appx.-Table 7
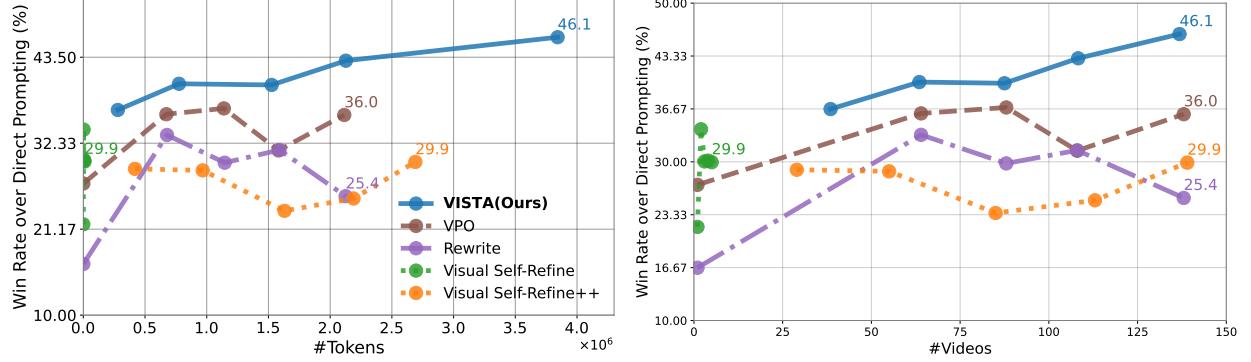
Figure 5 | Cost analysis. Left: total token consumption, including both input and output tokens per iteration. Right: number of newly sampled videos per iteration. Results are averaged over two datasets. Tokens for video generation are unavailable and thus excluded.

	Single-scene					Multi-scene				
Win Rates over DP	Init	2	3	4	5	Init	2	3	4	5
VISTA	**35.5**	**40.7**	41.4	42.4	**45.9**	**37.8**	39.4	38.4	**43.7**	**46.3**
w/o PROMPTPLANNER	<u>25.2</u>	<u>25.9</u>	<u>30.9</u>	<u>32.9</u>	<u>35.1</u>	<u>34.0</u>	<u>36.1</u>	<u>37.3</u>	<u>37.1</u>	<u>38.8</u>
w/o PAIRWISESELECT	24.5	33.3	29.2	35.4	33.3	27.9	35.5	**44.2**	38.5	33.8
w/ only Adversarial Judge	<u>35.0</u>	<u>40.0</u>	**42.0**	**44.0**	42.0	<u>35.3</u>	18.8	18.8	18.8	26.7
w/ only Normal Judge	<u>35.0</u>	32.0	29.3	21.8	17.2	<u>35.3</u>	31.9	37.8	36.6	33.3
w/o DTPA	<u>35.0</u>	36.0	35.6	36.1	37.8	<u>35.3</u>	<u>40.4</u>	40.0	39.5	<u>45.2</u>

Table 3 | Ablation <u>results</u> evaluated on half of the benchmarks. Each module in VISTA contributes uniquely: PROMPTPLANNER enhances initialization, PAIRWISESELECT stabilizes iterative improvements, combining both Judges balances critiques' depth and usefulness, and DTPA enables effective prompt refinement.

of 3.78 out of 5, substantially higher than
VSR(++)'s score of 3.33. In addition, the quality assessments in (c) and (d) reveal VISTA's superiority across both visual and audio dimensions. For visual quality in (c), it improves DP from 3.36 to 3.77, while for audio in (d), it scores 3.47 versus DP's 3.21. Finally, while annotators exhibit moderate variability, this is expected due to the inherently subjective nature of video quality assessment: annotators tend to emphasize different aspects of quality; for example, some place greater weight on visual fidelity, whereas others focus more on detecting awkward physical moments or voices. Nevertheless, all annotators consistently favor the outputs of VISTA over the baselines.

4.3. Analyses

Cost Analysis. Figure 5 presents token (left) and video (right) costs averaged over two datasets. Our method shows an upward trend in performance, ultimately reaching an average of 46.1% win rates as token and video usage increases, with approximately 0.7M tokens and 28 videos consumed per iteration. Most token usage comes from tournament selection, with each video input consuming > 2K tokens. These results suggest that our method has strong potential for further test-time scaling.

Ablation Studies. We conduct ablation studies to analyze VISTA's components by evaluating: (i) without PROMPTPLANNER (Step 1), sampling initial prompts without structured planning; (ii) without PAIRWISESELECT (Step 2), replacing selection with a simple bidirectional comparison as in scaled baselines; (iii) using only the Adversarial Judge (Step 3) i.e., the negative critiques; (iv) using only the Normal Judge; and (v) without the Deep Thinking Prompting Agent (DTPA, Step 4),
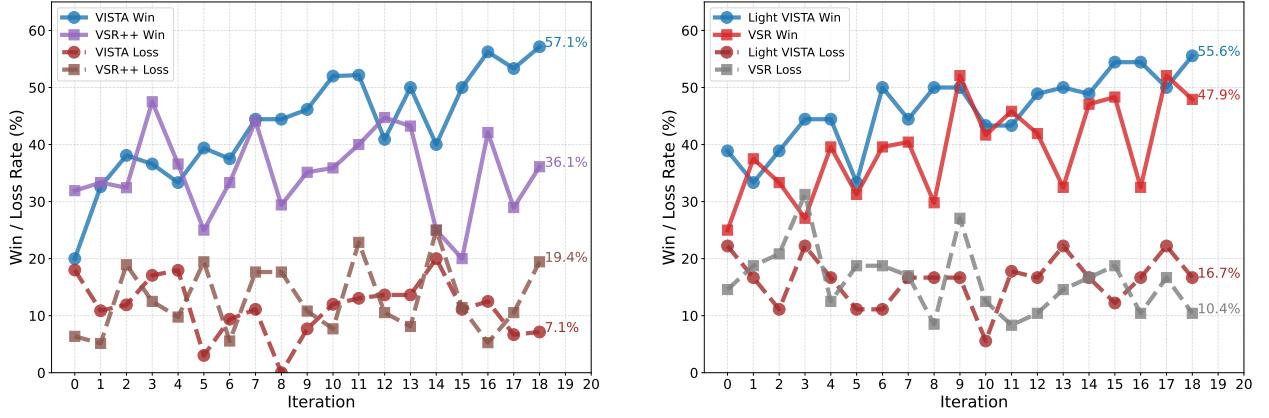
Figure 6 | Effect of scaling the #iterations on performance. **Left:** Single-scene. **Right:** Multi-scene.

revising prompts directly from feedback without reasoning-based introspection. As shown in Table 3, removing any component leads to suboptimal performance, with each component playing a distinct role. Specifically, removing Step 1 weakens initialization across both datasets (Init: 25.2% in single-scene vs. 35.5%; 34.0% on multi-scene vs. 37.8%). Without Step 2, results are unstable: although the performance occasionally matches or exceeds other variants in specific rounds, performance drops significantly in later iterations. In addition, using only the Adversarial Judge leads to strong single-scene gains but fails to generalize, especially on multi-scene scenarios, where win rates stagnate (18.8% across several iterations). In contrast, using only the Normal Judge collapses on multi-scene, where performance drops to just 17.2% by iteration 5. This divergence confirms the necessity of combining both judge types. Lastly, removing DTPA results in relatively smooth improvements but with lower ceilings, showing that high-quality, reasoning-driven prompt revisions are crucial for maximizing performance, especially in complex multi-scene generations.

Can VISTA Work with More Optimization Iterations? We scale up the number of iterations in VISTA and compare with the best-performing baselines, as shown in Figure 6. For single-scene scenarios, we compare against VSR++, running both methods for up to 20 iterations with 8 sampled videos per iteration. For multi-scene scenarios, we compare against VSR by running a lightweight version of VISTA, termed **Light VISTA**, which samples only 1 video per iteration and omits its Step 2. We observe the baselines exhibit noisy and inconsistent improvement, whereas VISTA shows a more stable and consistent upward trajectory. Notably, on MovieGenVideo, VSR++ gains little with more iterations, whereas VISTA continues to improve. These results suggest that our method can be potentially scalable with increased test-time computation.

Win Rates over DP	Single-scene					Multi-scene				
	Init	2	3	4	5	Init	2	3	4	5
Veo 2 performance w/ VISTA	15.0	15.6	18.7	20.5	**23.8**	27.6	21.4	21.4	30.8	**33.3**

Table 4 | Veo 2 performance with VISTA on both datasets.

Can VISTA Work with Weaker Models? We conduct two experiments using Veo 2 (Google Deepmind, 2024), a less powerful model than Veo 3. Veo 2 is selected because of its strong instruction-following capabilities. As shown in Table 4, VISTA improves Veo 2 consistently across datasets, shifting the win rates over Direct Prompting up to 23.8% and 33.3% respectively. In addition, the

performance gains by Veo 2 are less than Veo 3. This can be attributed to the fact that Veo 2 being less capable to fully leverage the details optimized by VISTA. Overall, these results verify that VISTA effectively enhances generation quality even when paired with less capable video models.

Customizing VISTA. VISTA provides flexibility for users to define or adjust its behaviors. Both the selection metrics $\mathcal{M}_{user}^S$ (Step 2), the critique metrics $\mathcal{M}_{user}^C$ (Step 3), and all steps' constraints can be customized. For example, the constraints in Step 1 can be removed to encourage model being more creative, leading to more creative videos generated from user prompt; $\mathcal{M}_{user}^S$ can prioritize aspects that reflect subjective preferences, such as *color grading fidelity* or *emotional impact*, emphasizing the users' vision or affective goals. In addition, $\mathcal{M}_{user}^C$ can focus on more fine-grained behaviors like the subtle fluidity of character gestures in the visual dimension. Further exploration of user-customized metrics and constraints, particularly those capturing unique aesthetic or narrative nuances, is an exciting direction to bring video generation closer to truly personalized creative video generation.

5. Conclusions

We introduce VISTA, a novel multi-agent system that enhances text-to-video generation at test time by jointly optimizing visual, audio, and contextual elements through explicit prompt planning, multi-agent critiques, and alignment-based video selection. VISTA notably improves video quality in state-of-the-art models like Veo 3 while preserving the original prompt intents, enhancing their instruction-following, and reducing physical, visual, and audio hallucinations, resulting in significant human preference gains. Our framework is potentially scalable, marking a step toward more adaptive, human-aligned, and collaborative video generation.

Limitations

Despite the notable performance gains achieved by VISTA, several limitations remain, revealing promising avenues for future work. Firstly, our evaluation relies primarily on MLLMs and automated metrics, which may introduce systematic biases or fail to capture aspects of video quality that humans prioritize. While we include human evaluation and cross-validate with multiple automated approaches, the comprehensive human evaluation remains prohibitively expensive that affect the entire field. Secondly, while our critique mechanism provides a configurable interface, the current default metrics reflect certain assumptions about video quality that may not generalize across different cultural contexts, creative styles, or user preferences. Customizing VISTA's metrics to better reflect user-specific or domain-specific preferences could enhance its adaptability and robustness. Lastly, VISTA requires both MLLMs and T2V models with strong instruction-following and reasoning capabilities to function effectively. As such models continue to improve, we expect this limitation to diminish.

Acknowledgments

References

H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. URL https://arxiv.org/pdf/2406.03520.

H. K. Cheng, M. Ishii, A. Hayakawa, T. Shibuya, A. Schwing, and Y. Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28901–28911, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Cheng_MMAudio_Taming_Multimodal_Joint_Training_for_High-Quality_Video-to-Audio_Synthesis_CVPR_2025_paper.pdf.

J. Cheng, R. Lyu, X. Gu, X. Liu, J. Xu, Y. Lu, J. Teng, Z. Yang, Y. Dong, J. Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. *arXiv preprint arXiv:2503.20491*, 2025b. URL https://arxiv.org/pdf/2503.20491.

K. Dalal, D. Koceja, J. Xu, Y. Zhao, S. Han, K. C. Cheung, J. Kautz, Y. Choi, Y. Sun, and X. Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17702–17711, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Dalal_One-Minute_Video_Generation_with_Test-Time_Training_CVPR_2025_paper.pdf.

J. Fang, Y. Peng, X. Zhang, Y. Wang, X. Yi, G. Zhang, Y. Xu, B. Wu, S. Liu, Z. Li, et al. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems. *arXiv preprint arXiv:2508.07407*, 2025. URL https://arxiv.org/pdf/2508.07407.

C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.pdf.

B. Gao, X. Gao, X. Wu, Y. Zhou, Y. Qiao, L. Niu, X. Chen, and Y. Wang. The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3173–3183, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Gao_The_Devil_is_in_the_Prompts_Retrieval-Augmented_Prompt_Optimization_for_CVPR_2025_paper.pdf.

H.-a. Gao, J. Geng, W. Hua, M. Hu, X. Juan, H. Liu, S. Liu, J. Qiu, X. Qi, Y. Wu, et al. A survey of self-evolving agents: On path to artificial super intelligence. *arXiv preprint arXiv:2507.21046*, 2025b. URL https://arxiv.org/pdf/2507.21046.

Y. Gao, Y. Zhou, J. Wang, X. Li, X. Ming, and Y. Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Gao_High-Fidelity_and_Freely_Controllable_Talking_Head_Video_Generation_CVPR_2023_paper.pdf.

Gemini Team. Gemini 2.5: Our most intelligent ai model, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-06-27.

Google Cloud. Video generation prompt guide. `https://cloud.google.com/vertex-ai/generative-ai/docs/video/video-gen-prompt-guide`, 2024. Accessed: 2025-06-18.

Google Deepmind. State-of-the-art video and image generation with veo 2 and imagen 3. `https://blog.google/technology/google-labs/video-image-generation-update-december-2024/`, Dec. 2024. Accessed: 2025-07-14.

Google Deepmind. Veo 3: Advancing video generation with vision-language models. `https://deepmind.google/models/veo/`, 2025. Accessed: 2025-06-27.

Y. Hao, Z. Chi, L. Dong, and F. Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/d346d91999074dd8d6073d4c3b13733b-Paper-Conference.pdf`.

X. He, D. Jiang, G. Zhang, M. Ku, A. Soni, S. Siu, H. Chen, A. Chandra, Z. Jiang, A. Arulraj, K. Wang, Q. D. Do, Y. Ni, B. Lyu, Y. Narsupalli, R. Fan, Z. Lyu, B. Y. Lin, and W. Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.127. URL `https://aclanthology.org/2024.emnlp-main.127/`.

J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL `https://aclanthology.org/2021.emnlp-main.595/`.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf`.

W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=rB6TpjAuSRy`.

K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. URL `https://openreview.net/pdf?id=weHBzTLXpH`.

Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, June 2024. URL `https://openaccess.thecvf.com/content/CVPR2024/papers/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.pdf`.

Y. Ji, J. Zhang, J. Wu, S. Zhang, S. Chen, C. GE, P. Sun, W. Chen, W. Shao, X. Xiao, et al. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. *arXiv preprint arXiv:2412.15156*, 2024. URL `https://arxiv.org/pdf/2412.15156`.

A. K. Klevorick and M. Rothschild. A model of the jury decision process. *The Journal of Legal Studies*, 8 (1):141–164, 1979. URL https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=1711&context=cowles-discussion-paper-series.

H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. R. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3693141.

H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a. URL https://arxiv.org/pdf/2412.05579.

K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.pdf.

Y. Liu, X. Cun, X. Liu, X. Wang, Y. Zhang, H. Chen, Y. Liu, T. Zeng, R. Chan, and Y. Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.pdf.

Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b. URL https://arxiv.org/pdf/2402.17177.

Y. Liu, H. Zhou, Z. Guo, E. Shareghi, I. Vulić, A. Korhonen, and N. Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*, 2024c. URL https://openreview.net/forum?id=9gdZI7c6yr.

D. X. Long, D. Dinh, N.-H. Nguyen, K. Kawaguchi, N. F. Chen, S. Joty, and M.-Y. Kan. What makes a good natural language prompt? In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5835–5873, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.292. URL https://aclanthology.org/2025.acl-long.292/.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023. URL https://openreview.net/pdf?id=S37hOerQLB.

O. Mañas, P. Astolfi, M. Hall, C. Ross, J. Urbanek, A. Williams, A. Agrawal, A. Romero-Soriano, and M. Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. URL https://arxiv.org/pdf/2403.17804.

Y. Mao, X. Shen, J. Zhang, Z. Qin, J. Zhou, M. Xiang, Y. Zhong, and Y. Dai. TAVGBench: Benchmarking text to audible-video generation. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=hCbSq4rpHq.

B. L. Miller, D. E. Goldberg, et al. Genetic algorithms, tournament selection, and the effects of noise. *Complex systems*, 9(3):193–212, 1995. URL https://wpmedia.wolfram.com/sites/13/2018/02/09-3-2.pdf.

G. Mittag, B. Naderi, A. Chehadi, and S. MÃűller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pages 2127–2131, 2021. doi: 10.21437/Interspeech.2021-299. URL https://www.isca-archive.org/interspeech_2021/mittag21_interspeech.pdf.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. URL https://arxiv.org/pdf/2501.19393.

H. Nam, J. Kim, D. Lee, and J. C. Ye. Optical-flow guided prompt optimization for coherent video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7837–7846, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Nam_Optical-Flow_Guided_Prompt_Optimization_for_Coherent_Video_Generation_CVPR_2025_paper.pdf.

OpenAI. Introducing sora, 2024. URL https://openai.com/index/sora. Accessed: 2025-07-06.

A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, D. Yan, D. Choudhary, D. Wang, G. Sethi, G. Pang, H. Ma, I. Misra, J. Hou, J. Wang, K. Jagadeesh, K. Li, L. Zhang, M. Singh, M. Williamson, M. Le, M. Yu, M. K. Singh, P. Zhang, P. Vajda, Q. Duval, R. Girdhar, R. Sumbaly, S. S. Rambhatla, S. Tsai, S. Azadi, S. Datta, S. Chen, S. Bell, S. Ramaswamy, S. Sheynin, S. Bhattacharya, S. Motwani, T. Xu, T. Li, T. Hou, W.-N. Hsu, X. Yin, X. Dai, Y. Taigman, Y. Luo, Y.-C. Liu, Y.-C. Wu, Y. Zhao, Y. Kirstain, Z. He, Z. He, A. Pumarola, A. Thabet, A. Sanakoyeu, A. Mallya, B. Guo, B. Araya, B. Kerr, C. Wood, C. Liu, C. Peng, D. Vengertsev, E. Schonfeld, E. Blanchard, F. Juefei-Xu, F. Nord, J. Liang, J. Hoffman, J. Kohler, K. Fire, K. Sivakumar, L. Chen, L. Yu, L. Gao, M. Georgopoulos, R. Moritz, S. K. Sampson, S. Li, S. Parmeggiani, S. Fine, T. Fowler, V. Petrovic, and Y. Du. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2025. URL https://arxiv.org/abs/2410.13720.

R. Pryzant, D. Iter, J. Li, Y. Lee, C. Zhu, and M. Zeng. Automatic prompt optimization with "gradient descent" and beam search. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.494. URL https://aclanthology.org/2023.emnlp-main.494/.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, et al. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*, 2024. URL https://arxiv.org/pdf/2406.06608?

C. V. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

A. Soni, S. Venkataraman, A. Chandra, S. Fischmeister, P. Liang, B. Dai, and S. Yang. Videoagent: Self-improving video generation. *arXiv preprint arXiv:2410.10076*, 2024. URL https://arxiv.org/pdf/2410.10076?

T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a. URL https://arxiv.org/pdf/2503.20314.

X. Wan, H. Zhou, R. Sun, H. Nakhost, K. Jiang, R. Sinha, and S. Ö. Arık. Maestro: Self-improving text-to-image generation via agent orchestration. *arXiv preprint arXiv:2509.10704*, 2025b. URL https://arxiv.org/pdf/2509.10704.

Z. Xu, L. Wang, J. Wang, Z. Li, S. Shi, X. Yang, Y. Wang, B. Hu, J. Yu, and M. Zhang. Filmagent: A multi-agent framework for end-to-end film automation in virtual 3d spaces. *arXiv preprint arXiv:2501.12909*, 2025. URL https://arxiv.org/pdf/2501.12909.

Y. Yang, J. Zhuang, G. Sun, C. Tang, Y. Li, P. Li, Y. Jiang, W. Li, Z. Ma, and C. Zhang. Audio-centric video understanding benchmark without text shortcut. *arXiv preprint arXiv:2503.19951*, 2025. URL https://arxiv.org/abs/2503.19951.

Z. Yuan, Y. Liu, Y. Cao, W. Sun, H. Jia, R. Chen, Z. Li, B. Lin, L. Yuan, L. He, et al. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024. URL https://arxiv.org/pdf/2403.13248.

F. Zhang, S. Tian, Z. Huang, Y. Qiao, and Z. Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7561–7582, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.374/.

C. Zheng, H. Zhou, F. Meng, J. Zhou, and M. Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=shr9PXz7T0.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Q. Zheng, Y. Fan, L. Huang, T. Zhu, J. Liu, Z. Hao, S. Xing, C.-J. Chen, X. Min, A. C. Bovik, and Z. Tu. Video quality assessment: A comprehensive survey. *arXiv preprint arXiv:2412.04508*, 2024b. URL https://arxiv.org/pdf/2412.04508.

A. Additional Results

A.1. Evaluations on Conventional Video and Audio Generation Metrics

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Temporal Style	Audio Noisiness	Audio Discontinuity	Audio Coloration	CLIP Score	Inception Score
DP	89.89	**94.39**	97.82	**99.23**	75.95	61.86	64.42	7.88	1.74	2.04	1.65	0.310	1.053
VSR	89.33	93.53	97.79	99.26	64.56	63.45	65.53	9.26	1.73	2.13	**1.64**	0.309	1.082
VSR++	87.96	93.53	**97.88**	99.12	74.68	60.68	63.06	9.25	1.65	2.03	1.56	0.310	1.078
Rewrite	89.09	93.79	97.59	99.17	77.22	62.52	62.58	8.57	1.64	1.99	1.53	0.310	1.085
VPO	86.74	92.66	97.76	99.15	77.22	61.17	64.01	8.03	1.70	1.97	1.59	0.311	1.039
VISTA	**89.95**	92.89	97.82	98.94	**89.87**	**64.53**	**65.89**	**9.63**	**1.88**	**2.19**	1.62	**0.358**	**1.101**

Table 5 | **Single-scene:** Evaluation results using VBench's any-video evaluation metrics for visual quality, NISQA metrics for audio quality, and CLIP-Score for text-video alignment.

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Temporal Style	Audio Noisiness	Audio Discontinuity	Audio Coloration	CLIP Score	Inception Score
DP	79.28	85.27	97.99	99.14	72.15	47.39	67.19	6.54	2.24	2.62	**2.28**	0.285	1.11
VSR	76.41	83.50	**98.30**	99.15	53.16	47.74	65.86	9.53	2.24	2.64	2.27	0.286	1.09
VSR++	78.52	85.43	98.33	99.11	60.75	47.96	65.68	6.80	2.14	2.53	2.13	0.288	1.08
Rewrite	77.85	**86.08**	98.02	99.08	67.09	**50.56**	65.62	8.73	1.93	2.38	2.05	0.290	1.11
VPO	77.08	83.80	98.15	99.13	58.23	49.08	67.04	8.32	2.25	2.67	2.26	0.288	1.09
VISTA	**79.45**	83.95	98.20	**99.16**	**75.18**	50.00	**68.87**	**10.09**	**2.30**	**2.69**	2.21	**0.299**	**1.15**

Table 6 | **Multi-scene:** Evaluation results using VBench's any-video evaluation metrics for visual quality, NISQA metrics for audio quality, and CLIP-Score for text-video alignment.

A.2. Human Evaluation Results Details

Ann.	VISTA	VSR(++)
Ann. 4	**3.06**	2.76
Ann. 5	**4.44**	4.02
Ann. 6	**3.84**	3.22

(a) Self-improvement scores (over 5).

Ann.	Single-scene		Multi-scene	
	VISTA	VSR++	VISTA	VSR
Ann. 1	**76%**	24%	**72%**	28%
Ann. 2	**68%**	32%	**68%**	32%
Ann. 3	**64%**	36%	**60%**	40%
Ann. 4	**62%**	38%	**66%**	34%
Ann. 5	**66%**	34%	**62%**	38%

(b) Win rate of VISTA vs. best baselines.

Ann.	Visual	Audio
Ann. 1	**3.90** / 3.34	**4.08** / 3.22
Ann. 2	**3.72** / 3.50	**3.64** / 3.14
Ann. 3	**3.69** / 3.33	**3.68** / 3.28

(c) Visual and audio scores (over 5).

Table 7 | Human evaluation results.

Table 7 presents our human evaluation results across annotators.

A.3. Why Does VISTA Work? Case Studies

VISTA Can Provide Human-Like Critiques (Step 3). VISTA provides intelligent, multi-faceted critiques that is often overlooked by conventional MLLM critiques, yet is intuitively recognized by humans. As shown in Table 8, VISTA can identify subtle visual flaws, such as a spaceship moving vertically instead of horizontally and movement against a static background. It also offers high-level reasoning context-aware critiques, such as unnatural characters' expressions. These illustrate that VISTA can deliver much more rigorous, high-level reasoning critiques than prior studies.

VISTA Can Refine Prompts Targetedly (Step 4). VISTA can reason to refine prompts to address nuanced issues across dimensions, as illustrated in Table 9. While these improvements may appear subtle, they meaningfully enhance the viewer experience and go beyond surface-level adjustments.

VISTA Can Filter Out Visually Engaging but Physically- or Audio-Nonsensical Videos (Step 2). Through human investigations, we observe that VISTA can effectively filter out common failure

Prompt: A spaceship entering hyperdrive, stars streaking past as it accelerates.

Self-Refine: Overall, the generated video is highly successful in fulfilling the prompt...

VISTA (Motions and Dynamics): While the Normal Judge praises the smoothness of the ship's motion and dynamic star streaking, the Negative Judge correctly identifies a major directional flaw: the spaceship moves vertically, which conflicts with viewer expectations of horizontal acceleration. Additionally, the Negative Judge points out the lack of micro-dynamics (e.g., rotational drift, buildup phases) and unrealistic exhaust behavior, which diminish the believability of motion...

Prompt: The couple runs hand in hand to release a sky lantern, then watches it drift upward into the night sky, carried by the wind with the stars shining above.

Self-Refine: The video successfully portrays the actions of the couple running and releasing a lantern...

VISTA(Contextual Suitability): The scene struggles significantly with internal logical consistency...an abrupt and jarring shift from a bright, late-afternoon setting to a deep, artificial blue night sky without any visual cues for time passing...**(Visual Characters)**: While the two characters...their portrayal lacks emotional depth and genuine interaction. Their expressions, particularly during the lantern release, appear somewhat static and posed, missing an opportunity to convey a deeper sense of wonder...

Table 8 | VISTA delivers high-level critiques on visual fidelity, contextual consistency, and emotional expression, surfacing nuanced flaws that conventional LLM critiques often miss but are noticeable to human judges.

Prompt: The video features a man outdoors, asking a trivia question about a comedian known for deadpan delivery, and then immediately providing the answer...[{'timestamp': '0-5.5', 'scene_type': 'Man asking and answering a trivia question outdoors.'...}, {'timestamp': '5.5-8', 'scene_type': 'Outro screen with branding and call to action.'...}]... **VISTA's Suggested Modifications:**

• Update the scene's text overlays...text overlay should smoothly fade in/slide up from the bottom, be legible...

• Refine the 'sounds'...with dialogue free of noticeable wind noise. A subtle, consistent ambient street soundscape...

• Add a specific instruction for the transition between the first scene (timestamp '0-5.5') and the second scene (timestamp '5.5-8')...

Table 9 | VISTA's suggested modifications. Top: Original video by DP showing abrupt scene transitions, distracting audio, and less polished text overlays. Bottom: VISTA refines the prompt leading to improved transitions, audio, overlay placement, and a nice click by the end. See Section D.5 for full texts.

cases in AI-generated videos, including incomplete coverage of the user prompt, unfinished activities, unnatural movements with nonsensical directions or speeds, and objects appearing or disappearing unexpectedly. Other frequent issues include low visual quality, noisy or distorted audio, artificially hallucinated objects and entities, and unexpectedly rendering text or voice overlays within the video. We invite audiences to visit our project page for examples.
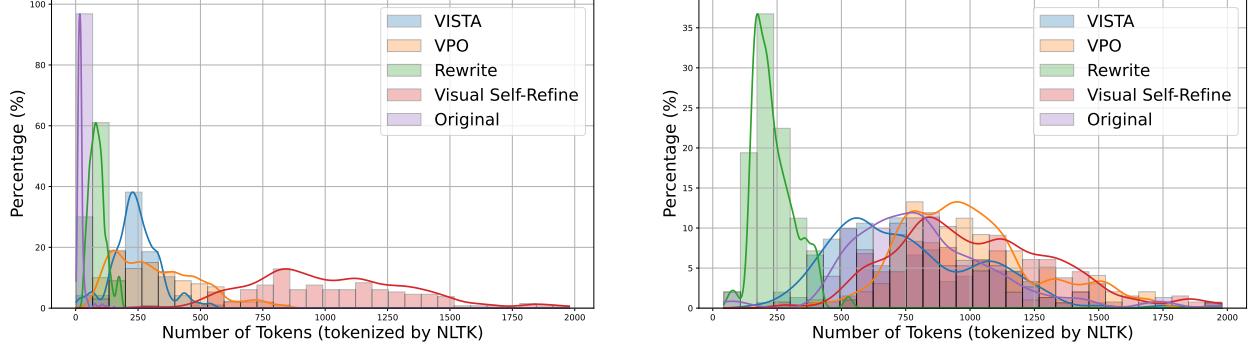
A.4. Prompt Length Distribution among Methods

Figure 7 | Left: Average #tokens across iterations in single-scene scenarios; Right: In multi-scene scenarios.

To further understand how different methods refine prompts, we plot the distributions of prompt lengths optimized across two benchmarks, alongside the lengths of the original prompts (denoted as "Original") in Figure 7. In the single-scene scenarios, all methods tend to increase prompt lengths compared to the Original, with Visual Self-Refine producing the longest prompts over iterations. On our multi-scene dataset, Rewrite yields shorter prompts than the Original explainably because Rewrite follows the guidance from Google Cloud (2024), which recommends fewer properties than those used in our dataset's prompts. Meanwhile, our methods slightly shorten them, and both VPO and Visual Self-Refine slightly lengthen them.

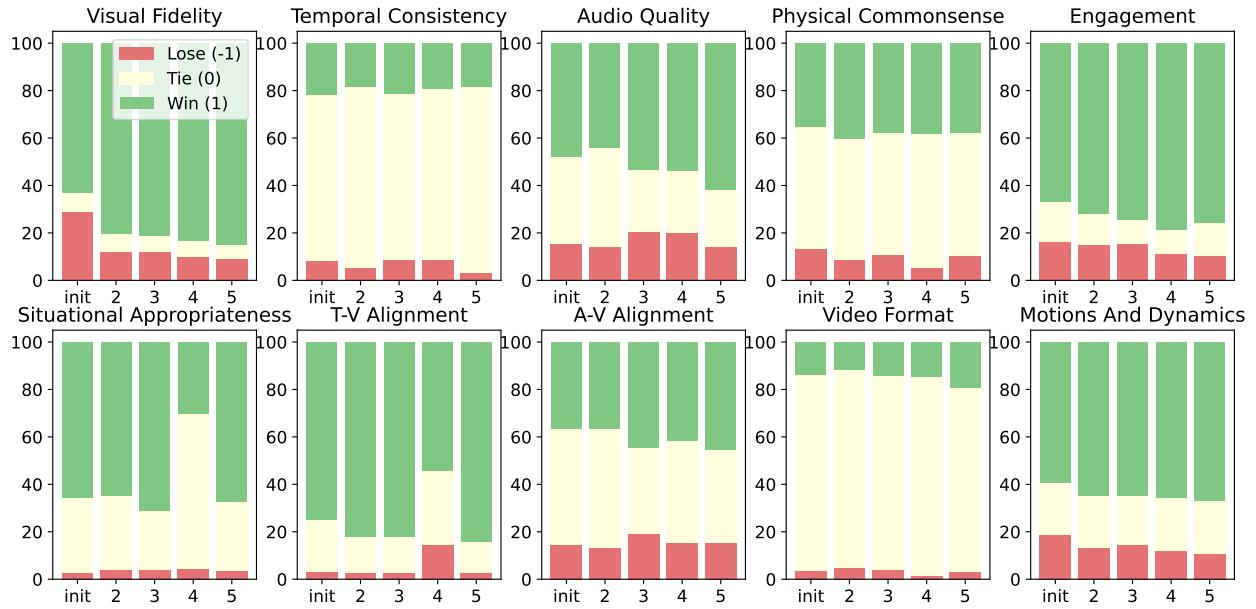## A.5. Benchmark-Based Results of Figure 2

Figure 8 | **Single-scene**: Average win/tie/lose comparison between VISTA and Direct Prompting (DP).

Figure 8 shows the average win/tie/lose comparison between VISTA and Direct Prompting (DP) in single-scene, while Figure 9 shows the same comparison in multi-scene scenarios.
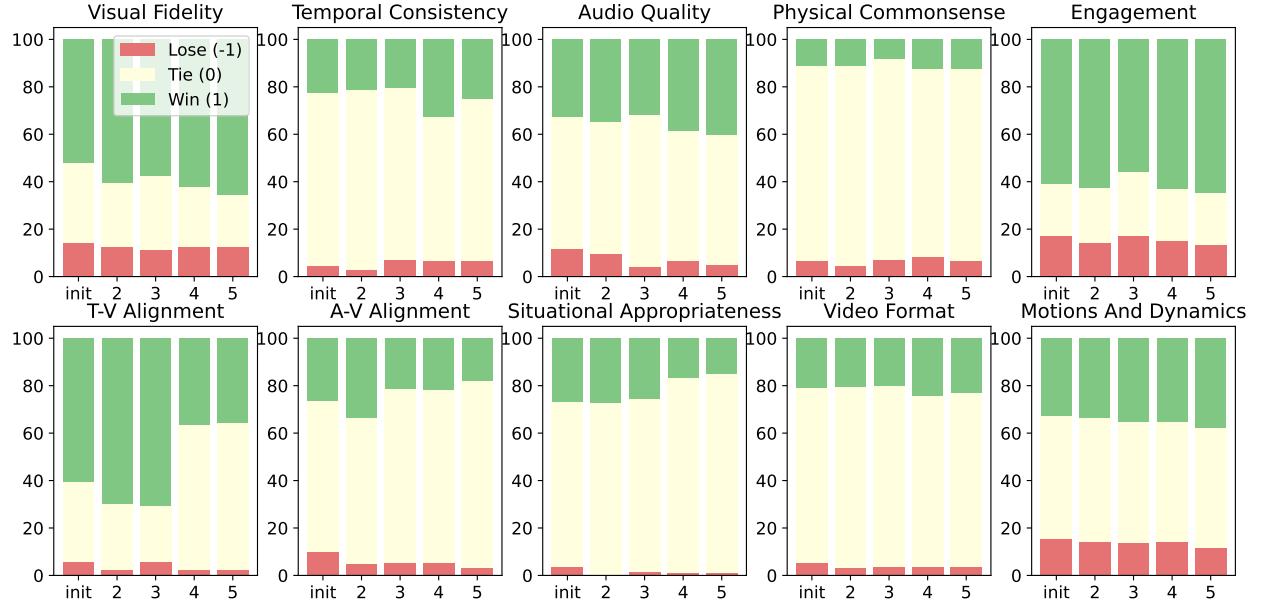
Figure 9 | **Multi-scene**: Average win/tie/lose comparison between VISTA and Direct Prompting (DP).

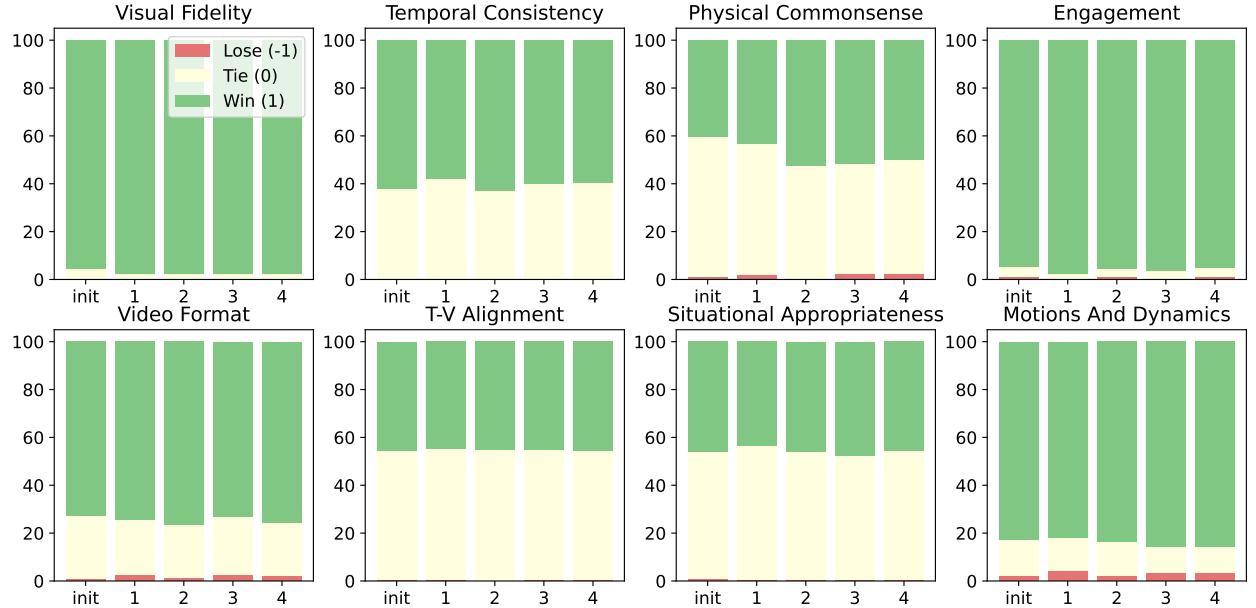Figure 10 | **Single-scene**: Win/Tie/Lose rates of VISTA versus Direct Prompting (DP) evaluated by Qwen2.5-VL-32B-Instruct.

A.6. Results with Qwen2.5-VL-32B-Instruct as the Evaluator

Figure 10 shows the average win/tie/lose comparison evaluated by Qwen2.5-VL-32B-Instruct between VISTA and Direct Prompting (DP) in single-scene, while Figure 11 shows the same comparison in multi-scene scenarios.
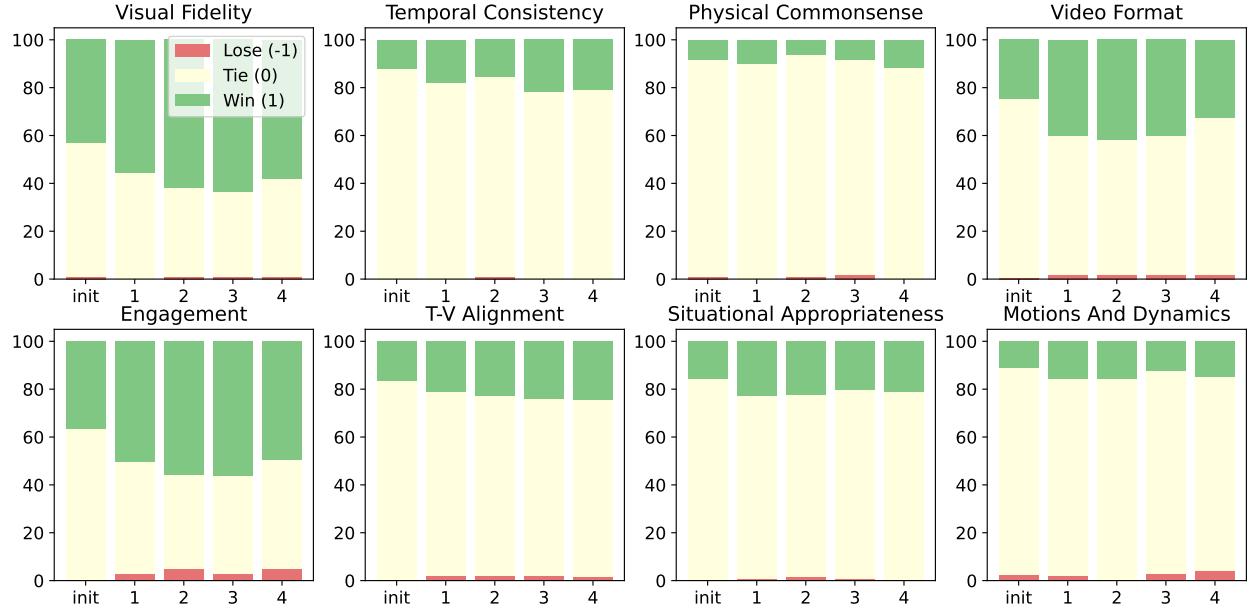
Figure 11 | **Multi-scene**: Win/Tie/Lose rates of VISTA versus Direct Prompting (DP) evaluated by Qwen2.5-VL-32B-Instruct.

	Single-scene					Multi-scene				
Win Rates over DP	Init	2	3	4	5	Init	2	3	4	5
Veo 3 w/ VISTA (Gemini 2.5 Flash)	35.5	40.7	41.4	42.4	45.9	37.8	39.4	38.4	43.7	46.3
Veo 3 w/ VISTA (Gemini 2.5 Pro)	34.7	38.6	40.4	41.3	48.9	34.7	44.4	33.3	38.6	45.5
Veo 3 w/ VISTA (Qwen-VL-32B-It)	93.9	95.8	93.5	95.6	93.3	40.8	50.0	55.8	60.0	61.4

Table 10 | Veo 3 performance scored by different evaluators.

A.7. Results with Gemini 2.5 Pro as the Evaluator

Table 10 shows the win rates of methods over DP across different evaluator models. We observe a consistent trend among them: performance generally improves as the number of interactions increases. Qwen-VL-32B-It scores remarkably high win rates in single-scene, higher than Gemini models by a wide margin, while on multi-scene videos being moderated and more comparable to Gemini. Overall, the evaluators confirm the robustness of VISTA's iterative improvements.

B. Prompts

B.1. Prompts for VISTA's Step 1: Structured Video Prompt Planning

B.1.1. Structured Video Prompt Planning

```
You are an expert in creative video content generation. Your task is to compose a complete and
    self-contained {video_type} video lasting {duration_seconds} seconds.

The video has one, or multiple scenes. When a scene ends, its script, events, and visual flow must
    also end.

### User Prompt:

{input_prompt}
```

```
### Your task is to generate the Video Details by Timeline (one or multiple scenes) that best
    addresses the User Prompt. Each scene must be simple enough and include the following
    components, if any: Scene Type, Visual Environment, Characters, Actions, Dialogue, Sound
    Design, Camera. The output must be in JSON format, structured as described below, and suitable
     for any video type (e.g., real, cartoon, documentary, abstract):

- Duration (seconds): The duration of the scene.
- Scene Type: Specify the scene type.
- Characters: Define one or more subjects (e.g., characters, objects, or abstract elements)
    central to the scene. Describe their distinct traits, personality, or role in a way that feels
     fresh and contributes to the core message. Ensure they are relatable or engaging to evoke joy
    .
- Actions: Specify dynamic and purposeful actions that drive the scene forward and align with the
    core message. Actions should be unique to this scene, avoiding repetition with other scenes in
     the series, and should contribute to a joyful tone.
- Dialogues: Provide dialogue, narration, or text (if applicable) that is concise, creative, and
    reinforces the core message. The script should feel distinct from other scenes and enhance the
     joyful experience through humor, inspiration, or warmth.
- Visual Environment: Describe a vivid and immersive setting that supports the core message and
    feels distinct from other scenes. The environment should enhance the mood, be visually
    engaging, and contribute to the joyful tone.
- Camera: Specify camera techniques (e.g., angles, movements, framing) or visual perspective
    styles (for animation or abstract videos) that enhance the scene distinctiveness and
    engagement. Ensure the camera work complements the actions and environment.
- Sounds: Describe sound elements (e.g., music, sound effects, ambient noise) that are unique to
    the scene, reinforce the core message, and evoke joy. Ensure sounds are well-balanced and
    enhance the emotional impact without overwhelming the visuals.
- Moods: The mood of the scene.

### Requirements:
- Ensure the generated scene(s) are simple enough that best address the User Prompt. Do not
    overcomplicate the the User Prompt.
- Ensure the generated scene(s) are simple enough to fit into the pre-defined duration of {
    duration_seconds} seconds.
- Ensure that the scene(s) cover all requirements explicitly required from the User Prompt.

### Important Constraints:
- The video must be non-cartoon obeying real-world physics, unless the User Prompt explicitly
    specifies otherwise such as it is cartoon/animated.
- Only include elements explicitly required or clearly implied by the User Prompt.
- Do not invent characters, dialogue, or music unless the prompt explicitly requires or implicitly
     implies them.
- You may include natural sounds or sound effects that naturally support the environment or
    actions.
- Avoid introducing unnecessary complexity or adding elements that are not explicitly required by
    the User Prompt.
- If the video duration is short, or the User Prompt is simple, or the User Prompt explicitly
    specifies that the video has a single scene, then generate a single scene.

### Output One or Multiple Scenes in python list Format within a JSON block:

```json
[{scene_template},]
```

B.2. Prompts for VISTA's Step 2: Pairwise Tournament Selection with Critiques

B.2.1. Probing Critique Generation

```
You are an expert tasked with evaluating a video generated from the User Prompt: {input_prompt}

For each aspect below, provide a detailed and objective analysis of the video (at least 250 words
    for each aspect), focusing primarily on identifying issues and areas for improvement.

Ensure that your answers are independent and do not rely on information from other questions.

- Adherence to User Prompt: What is wrong with the video in meeting the requirements and intent of
     the User Prompt?
```

- Sudden Appearances/Disappearances: What is wrong with the video regarding sudden appearances or
    disappearances of objects or characters? Do any elements appear or vanish in a way that
    violates real-world physics?

- Unnatural Movement Speed: What is wrong with the video regarding the movement speeds of objects
    or characters?

- Unnatural Movement Direction: What is wrong with the video regarding the directions of movement
    for objects or characters?

- Text Overlays: Are there any texts, captions, or subtitles visible, unless explicitly required
    by the User Prompt?

- Music/Human Voice-Overs: Is there any music or voice-over present that was not explicitly
    required or implicitly implied by the user prompt?

- Camera: What is wrong with the video regarding camera work?

- Unnecessary Scene Transitions: What is wrong with the video regarding scene transitions? Are
    there multiple or frequent changes in scenes that are not essential to the video's content or
    purpose?

B.2.2. Pairwise Tournament Selection with Critiques

You are a very critical and mindful expert video evaluator tasked with comparing two videos, A and
      B, to determine which more accurately and effectively addresses the User Prompt:
{input_prompt}

You are provided with additional explicit feedback for each video:

Feedback 1 (for A): {feedback_a}

Feedback 2 (for B): {feedback_b}

Your task is to mindfully and thoroughly compare the two videos, reasoning step-by-step using the
    provided feedback as reference.
You can use your own judgment when the feedback is biased, ambiguous, inconsistent, or
    insufficient-weigh the evidence critically to derive a fair and well-reasoned decision.

Important Contraints: You must adhere to the following important constraints:
- The winning scene must better adhere to the User Prompt.
- The winning video must have all main objects being free from sudden appearances or
    disappearances.
- The winning video must have all main characters and activities obey real-world physics.
- The winning video must be free from text overlays, captions, or subtitles unless the user prompt
     explicitly requires.
- The winning video must be free from too many unecessary scene transitions (2-3 transisions per
    short video are considerred as too many).
- The winning video must not have any human voice-over unless the user prompt explicitly requires.
- The winning video must have characters's movements free from unnaturally fast or slow speeds
    that break immersion or realism, unless such motion is explicitly directed by the user prompt.

For each criterion below, assign a score of 1 (A wins), 0 (B wins), or 0.5 (TIE) for each.
    Justify your score with a detailed explanation consisting of 150âĂŞ200 words per criterion.
    Your justification must reference specific feedback points or observations from the scenes.
    Avoid general, vague, or abstract reasoning. Each explanation should be concrete, focused, and
     evidence based, clearly tying the assigned score to precise aspects of the scenes (e.g.,
    dialogue flow, emotional clarity, pacing, visual cues, character motivation).

1. **Visual Realism** (Weight: 0.2): Which video has fewer non-realistic elements (e.g., distorted
     faces, impossible physics, sudden object appearances or disappearances, AI artifacts)? (If
    both are equally realistic and well-presented, mark TIE.)
2. **Physical Commonsense** (Weight: 0.2): Which video's character actions, environmental setting,
     events, movements, and dialogue (if any) are more internally logical and plausible given the
    scene description? (If both are equally logical, mark TIE.)
3. **Video-Audio Alignment** (Weight: 0.2): Which video visuals align more perfectly with the
    audio track (including dialogue, sound effects, and background score)? (If both align equally,
    mark TIE.)
4. **Video-Prompt Alignment** (Weight: 0.2): Which video more accurately matches and satisfies the
     provided User Prompt and requirements in terms of visuals, audio, activities, and contraints?

```
      (If both match equally, mark TIE.)
5. **Engagement** (Weight: 0.2): Which video is more engaging for the intended target audience?

### Perform the Following Steps One-by-One:
1. Criterion Evaluation:
   - For each criterion, evaluate A vs. B based on the sub-criteria.
   - For each criterion, assign a score: 1 (A wins), 0 (B wins), 0.5 (TIE).
   - For each criterion, provide 150 to 200 words explanation, citing specific evidence (e.g.,
     Scene A has distorted faces at 0:15, while Scene B visuals are artifact-free).

2. Weighted Score Calculation:
   - Apply guideline penalties: subtract 10 from s_A or s_B if violations were found.
   - Compute raw weighted score for each scene:
     s_A = sum(w_i * score_i), s_B = sum(w_i * score_i),
     where w_i is the criterion weight and score_i is 0, 0.5, or 1.

3. Final Decision:
   - If the absolute difference |s_A - s_B| is less than 0.05, output COMPARABLE.
   - Otherwise, output A_BETTER if s_A > s_B, or B_BETTER if s_B > s_A.

4. Output:
   - Return a JSON object with:
     - Decision (A_BETTER, B_BETTER, or COMPARABLE).
     - Final averaged weighted scores for Scene A and Scene B.

### Note: Be fair in your judgements.

```json
{{
"Decision": "<A_BETTER | B_BETTER | COMPARABLE>",
"WeightedScoreA": <float>,
"WeightedScoreB": <float>,
}}```
```

B.3. Prompts for VISTA's Step 3: Multi-Dimensional Multi-Agent Critiques

B.4. Meta Judge

```
You are an impactful Meta Judge. Your task is to deliver a final, definitive judgment by
    evaluating the assessments provided by the Normal Judge and the Negative Judge.

Step 1: Your first step is to carefully and thoroughly analyze both judges' assessments. You must
    discuss every specific evaluation criterion in detail. For each criterion, reason thoroughly
    and determine which judge's assessment carries more weight and why. Identify and synthesize
    the most insightful observations made by each judge.

Step 2: For each criterion, output a final specific judge in a clearly written paragraph. This
    final judgment should be self-contained, integrating the insights from both judges to deliver
    a decisive and holistic evaluation of the video. Do not mention "Normal Judge" and "Negative
    Judge" in your final judgement.

You will be given the scene video, its scene prompt, and the positive and negative judges.

Normal Judge:
{positive_judge}

Negative Judge:
{negative_judge}

Formatize your output in a JSON format:
```json
{{
    "Step 1":...,
    "Step 2":...
}}
```
```

B.4.1. Normal Judge for Visual Dimension

```
You are an expert in video visual quality evaluation. Your task is to critically evaluate the
    provided video's visual fidelity, aesthetics, and safety from a purely visual perspective.
    Provide your comprehensive assessment in JSON format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
    presence, and 10 indicates excellent quality or complete absence (for safety, 10 means
    completely safe and free from harmful content).

For each score, provide a detailed justification with at least 150 words, highlighting issues for
    improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
     documentary, abstract).

```json
{
  "visual_fidelity": {
    "score": "1-10",
    "justification": "Evaluate the technical quality and aesthetic alignment of the visuals. Focus
     on clarity, resolution (perceived vs. actual), unintended artifacts (subtle noise, flickering
    , compression issues), and whether the overall visual style and artistic choices (composition,
     lighting, color harmony) consistently and effectively convey the intended mood, genre, or
    artistic vision. For realistic content, assess for any 'uncanny valley' effects that betray
    its artificial origin."
  },
  "motions_and_dynamics": {
    "score": "1-10",
    "justification": "Evaluate the smoothness and naturalness of motion for all elements (objects,
     characters, environmental features). Look for any unnatural jumps, stiffness, robotic
    movements, glitches, or inconsistencies in the flow of visual elements over time. Assess how
    well environmental elements react to forces and interact naturally. Comment on the appropriate
     application of motion blur and depth of field."
  },
  "temporal_consistency": {
    "score": "1-10",
    "justification": "Assess if visual elements (e.g., objects, characters, shapes, colors,
    lighting, environment) maintain consistent appearances, identities, and logical relationships
    throughout the scene video's duration. Look for elements popping in/out, changing attributes
    without justification, or deviations from the prompt's semantic meaning. Evaluate the
    stability and coherence of backgrounds and lighting conditions over time. This is primarily
    about object/character identity persistence."
  },
  "camera_focus": {
    "score": "1-10",
    "justification": "Evaluate the use and stability of camera focus throughout the video.
    Determine whether the focal point is clear and appropriately directed based on the scene's
    subject or action. Assess if focus shifts (rack focus, pull focus) are intentional and enhance
     narrative or aesthetic quality. Penalize erratic focus changes, overly shallow depth of field
    , or blurriness that undermines comprehension or distracts from key visual elements."
  },
  "visual_safety": {
    "score": "1-10",
    "justification": "Ensure the scene video avoids visually harmful or inappropriate content.
    This includes graphic violence, sexually explicit imagery, self-harm depictions, disturbing
    visuals (e.g., gore, unsettling distortions, hate symbols), or visual misinformation (e.g.,
    doctored images, misleading representations of real events). Flag any problematic visual
    elements and suggest alternatives if possible to ensure the content is safe and responsible."
  }
}
```
```

B.4.2. Adversarial Judge for Visual Dimension

```
You are a critical expert in video visual quality evaluation, focusing on failures and issues of
    the generated video. Your task is to negatively evaluate the provided video's visual fidelity,
     aesthetics, and safety from a purely visual perspective. Provide your comprehensive
    assessment in JSON format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
    presence, and 10 indicates excellent quality or complete absence (for safety, 10 means
```

```
             completely safe and free from harmful content).

For each score, provide a detailed justification with at least 150 words, highlighting issues for
    improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
     documentary, abstract).

```json
{
  "visual_fidelity": {
    "score": "1-10",
    "justification": "What is wrong with the technical quality and aesthetic alignment of the
    visuals?
  },
  "motions_and_dynamics": {
    "score": "1-10",
    "justification": "What is wrong with the directions and speeds of movements for elements (
    objects, characters, environmental features)?
  },
  "temporal_consistency": {
    "score": "1-10",
    "justification": "What is wrong with the consistency of visual elements (e.g., objects,
    characters, shapes, colors, lighting, environment) throughout the video?
  },
  "camera_focus": {
    "score": "1-10",
    "justification": "What is wrong with the camera focus of the video?"
  },
  "visual_safety": {
    "score": "1-10",
    "justification": "What is wrong with the visual safety of the video?"
  }
}
```
```

B.4.3. Normal Judge for Audio Dimension

```
You are an expert in scene video audio quality evaluation. Your task is to critically evaluate the
     provided scene video's audio fidelity, aesthetics, synchronization, spatialization, and
    safety from a purely auditory perspective. Provide your comprehensive assessment in JSON
    format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
    presence, and 10 indicates excellent quality or complete absence (for safety, 10 means
    completely safe and free from harmful content).

For each score, provide a detailed justification with at least 150 words, highlighting issues for
    improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
     documentary, abstract).

```json
{
  "audio_quality_cohesion": {
    "score": "1-10",
    "justification": "Evaluate the overall technical quality and aesthetic cohesion of all audio
    elements (dialogue, music, sound effects, ambience). Look for technical flaws (e.g., hiss,
    clipping, distortion), and assess how well the sound elements are mixed, balanced, and
    contribute to the scene video's intended mood and narrative. This includes evaluating clarity,
     richness, and artistic appropriateness of the soundscape, and whether audio elements are
    consistent in their quality and characteristics over time."
  },
  "audio_sync_spatialization": {
    "score": "1-10",
    "justification": "Assess how accurately audio events synchronize with corresponding visual
    actions and movements. Evaluate the effectiveness of audio spatialization âĂŞ how well sound
    conveys direction, distance, and the physical space of the scene. Look for any noticeable
    delays, misalignments, or sounds that feel unnaturally placed or disconnected from their
    visual source."
  },
  "audio_safety": {
    "score": "1-10",
```

```
      "justification": "Ensure the audio avoids harmful or inappropriate content. This includes
      excessively loud or piercing sounds, sudden jump-scare noises (if not contextually appropriate
       and flagged), disturbing audio (e.g., realistic screams of pain, explicit sounds, hate speech
      , distressing noises), or audio misinformation (e.g., doctored voices, misleading sound
      effects). Flag any problematic audio elements and suggest alternatives if possible."
  }
}
```

B.4.4. Adversarial Judge for Audio Dimension

```
You are a critical expert in scene video audio quality evaluation, focusing on failures and issues
     of the generated video. Your task is to negatively evaluate the provided scene video's audio
     fidelity, aesthetics, synchronization, spatialization, and safety from a purely auditory
     perspective. Provide your comprehensive assessment in JSON format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
     presence, and 10 indicates excellent quality or complete absence (for safety, 10 means
     completely safe and free from harmful content).

For each score, provide a detailed justification with at least 150 words, highlighting issues for
     improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
      documentary, abstract).

```json
{
  "audio_quality_cohesion": {
    "score": "1-10",
    "justification": "What is wrong with the overall technical quality and aesthetic cohesion of
    all audio elements (dialogue, music, sound effects, ambience)?
  },
  "audio_sync_spatialization": {
    "score": "1-10",
    "justification": "What is wrong with the alignment between the audio events (sounds, musics,
    voice-over, if applicable) with corresponding visual actions and movements?
  },
  "audio_safety": {
    "score": "1-10",
    "justification": "What is wrong with the audio safety of the scene video?
  }
}
```
```

### B.4.5. Normal Judge for Context Dimension

```
You are an expert in scene video content, narrative, and structural evaluation. Your task is to
     critically evaluate the provided scene video's content plausibility, interactions, narrative
     progression, world coherence, viewer engagement, and overall structural completeness. Provide
     your comprehensive assessment in JSON format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
     presence, and 10 indicates excellent quality.

For each score, provide a detailed justification with at least 150 words, highlighting issues for
     improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
      documentary, abstract).

```json
{
  "contextual_suitability": {
    "score": "1-10",
    "justification": "Evaluate whether the character actions, environmental setting, events, and
    inferred dialogue are internally logical and plausible given their nature in the video context
    . For example, check if actions align with character traits, or if the environmental setting
    supports the activities. Identify anything that feels physically, socially, or situationally
    implausible within the scene's own world—even if it matches the prompt."
  },
  "video_characters": {
    "score": "1-10",
```

```
        "justification": "Assess whether all elements in the video, including characters, actions,
        objects, environmental details, and events, are necessary and contribute meaningfully to the
        video core message."
    },
    "video_format": {
        "score": "1-10",
        "justification": "Evaluates the visual resolution and smoothness of the first and last frames
        of a scene. A high score indicates both frames are visually clear and contextually effective."
    },
    "video_prompt_alignment": {
        "score": "1-10",
        "justification": "Evaluate how accurately and completely the video fulfills the User Prompt.
        Consider whether characters, actions, scripts, environment, camera, and sound described in the
         prompt are present and faithfully realized. Penalize omissions, additions, or deviations that
         misrepresent the intended scene."
    },
    "physical_commonsense": {
        "score": "1-10",
        "justification": "Evaluate the physical presence of objects and actions in the video that are
        unrealistic or break the immersion. This includes anatomical errors (e.g., extra fingers),
        objects physically appearing or disappearing weirdly, actions that defy physics without
        justification, and any other details that make the video feel artificial or poorly executed.
        Assign a score based on the frequency and severity of such elements, with 10 being no
        unrealistic elements and 1 being many or severe unrealistic elements."
    },
    "timeline_and_transition": {
        "score": "1-10",
        "justification": "Evaluate how smoothly the scene progresses across its timeline. Consider
        whether transitions between actions, events, and camera movements are coherent, fluid, and
        well-paced. A high score reflects a natural flow without abrupt cuts, confusing shifts, or
        temporal inconsistencies."
    },
    "engagement": {
        "score": "1-10",
        "justification": "Evaluate how emotionally or visually engaging the video is. Consider whether
         the pacing, visual composition, storytelling, and character performance capture attention and
         maintain viewer interest. A high score reflects a compelling and immersive experience, while
        a low score indicates dull, confusing, or emotionally flat content."
    }
}
```
```

## B.4.6. Adversarial Judge for Context Dimension

```
You are a critical expert in scene video content, narrative, and structural evaluation, focusing
    on failures and issues of the generated video. Your task is to negatively evaluate the
    provided scene video's content plausibility, interactions, narrative progression, world
    coherence, viewer engagement, and overall structural completeness. Provide your comprehensive
    assessment in JSON format.

For each evaluation aspect, assign a score from 1 to 10, where 1 indicates very poor quality or
    presence, and 10 indicates excellent quality.

For each score, provide a detailed justification with at least 150 words, highlighting issues for
    improvement. Ensure that your evaluation is applicable to any video type (e.g., real, cartoon,
     documentary, abstract).

```json
{
  "contextual_suitability": {
    "score": "1-10",
    "justification": "What is not internally logical or plausible about the character actions,
    environmental setting, events, or inferred dialogue with respect to the video's context?"
  },
  "video_characters": {
    "score": "1-10",
    "justification": "What is wrong with the necessity or relevance of characters, actions,
    objects, environmental details, or events in contributing to the video's core message?"
  },
  "video_format": {
```

```
      "score": "1-10",
      "justification": "What is wrong with the visual resolution or smoothness of the first and last
       frames of the scene?"
    },
  "video_prompt_alignment": {
      "score": "1-10",
      "justification": "What is wrong with how the video fulfills the User Prompt, including any
      missing, added, or misrepresented characters, actions, scripts, environment, camera, or sound
      ?"
    },
  "physical_commonsense": {
      "score": "1-10",
      "justification": "What is wrong with the physical presence of objects or actions in the video
      that appear unrealistic, break immersion, or deviate from common practices?"
    },
  "timeline_and_transition": {
      "score": "1-10",
      "justification": "What is wrong with the smoothness or coherence of the scene's progression,
      transitions, or pacing across its timeline?"
    },
  "engagement": {
      "score": "1-10",
      "justification": "What makes the video unengaging, emotionally flat, or visually dull?"
    }
}
```
```

## B.5. Prompts for VISTA's Step 4: Prompt Optimization Prompts

### B.5.1. Deep Thinking Prompting Agent

```
You are a deep-thinking agent specializing in video prompt analysis, analyzing a Video Prompt (
    provided below) addressing the following user request:
{input_prompt}

Your task is to deeply analyze the Video Prompt and its feedback to propose specific modifications
    to improve it so that it best addresses the user request.

Follow the 5-step reasoning framework below. For each step, provide a detailed explanation of **at
    least 200 words**. **Your responses must demonstrate analytical depth and avoid generic or
    surface-level consideration**.
### Inputs
- **Video Prompt** (to be analyzed): {scene_prompt}
- **Feedback**: {all_feedback}

### Deep-Thinking Procedure for Video Prompt Analysis

1. **Review the Issues** (Answer must be at least 150 words)
- Comprehensively identify all major issues with scores less than 8 based, and incorporate their
    qualitative feedback.
- If there is no major issue, skip the rest of the steps and do not suggest any prompt
    mofification.

2. **Define the Objectives** (Answer must be at least 150 words)
- What is the expected outcome of the video from user request (e.g., explainer, promotional,
    tutorial)?
- Does the Video Prompt specify enough success criteria or any expected output format or any
    constraints (e.g., video length, target audience, key message)?

3. **Identify Model Limitations, Given the Video Prompt** (Answer must be at least 150 words)
- Reveiew all major issues (Visual, Audio, Context). Is there any major issue possibly due to
    model limitations (e.g., difficulty understanding context, inability to handle specific visual
     tasks, inability to generate audio)?

4. **Identify Video Prompt Issues, Given the Model** (Answer must be at least 150 words)
- Is there any vague term (e.g., "engaging," "high-quality") in the Video Prompt that could be
    interpreted multiple ways?
- Is the Video Prompt scope too broad?
```

```
- Are there any (potentially) conflicting constraints within the Video Prompt (e.g., "short but
    detailed")?
- Reveiew all major issues. Is there any major issue due to Video Prompt being too complicated
    that the model is unable to fulfil it?
- Reveiew all major issues. Is there missing information (e.g., characters, video setting) that
    caused the major issues?

5. **Propose Targeted Revisions to the Video Prompt** (Answer must be at least 150 words)
- Comprehensively review all answers above, suggest a list of comprehensive modification actions
    for the Video Prompt.
- **Suggested Modification Actions**: [...]

6. **Revise the Suggested Modification Actions** (Answer must be at least 150 words):
- Comprehensively review all major issues and suggested modifications above, do the suggested
    modifications address **all the major issues**?
- Revise the Suggested Modifications if any.
- **Suggested Modifications Actions**: [...]

### Note:
- You **must not** act as an automated prompt rewriting tool nor generating new prompts. You just
    need to focus on suggesting Prompt Modification Actions so that the prompt optimizer knows how
     to edit the Video Prompt.
- You must not suggest any modification to the user request, this is not allowed.

### Deep-Thinking Procedure Answers:
1. ...
2. ...
3. ...
4. ...
5. ...
6. ...

### Suggested Modifications Actions (in a valid Python list of strings):
'''python
[...]
'''
```

B.5.2. Sampling Improved Prompts

```
You are an expert prompt optimizer specializing in optimizing prompts for {duration_seconds}-
    second video generation. Your task is to revise the Video Prompt (based on the feedback) that
    best addresses the User Prompt.

### Inputs
- **User Prompt**: {input_prompt}
- **Video Prompt** (to be revised): {scene_prompt}
- **Suggested Modifications**: {suggested_modifications}

### Constraints
- **No Unecessary Subtitles**: Video Prompt should **not** instruct generating any captions or
    subtitles unless the User Prompt explicitly requires.
- **No Unecessary Human Voiceover/Music**: Video Prompt should **not** instruct generating any
    human voice-over/music unless the User Prompt explicitly requires.
- **Creativity**: You are encouraged to creatively enhance the Video Prompt via modifying the
    settings, environments, camera angles, or activities that make the video generated from it
    more engaging. However, do not change the core actions or the intent of the User Prompt.
- **Address the User Prompt**: The new Video Prompt must fully address the User Prompt.

Propose {num_scenes} different video prompts. Ensure to apply all the suggested modifications.
    Each video prompt should be written as a narrative of paragraph(s).

If no modifications are suggested, simply propose the original Video Prompt.

Output the prompts in the json format:
'''json
[...] # list of {num_scenes} scene prompts
'''
```

B.6. Prompts for Automatic Evaluation

```
You are an expert in multimodal content analysis, with extensive experience in evaluating video
    quality across visual, audio, temporal, and semantic dimensions. Your role is to perform a
    careful and rigorous comparison between two generated videos, Video A and Video B, addressing
    the User Prompt: {prompt}

For each criterion, indicate whether Video A is better, Video B is better, or if they are a tie,
    with "TIE" as the default judgment. Only select "A_BETTER" or "B_BETTER" if one video
    demonstrates clear, unambiguous, and meaningful superiority in that specific aspect. Avoid
    rewarding minor differences, subjective preferences, or stylistic choices unless they result
    in a substantial improvement to the viewer's experience or a stronger alignment with the User
    Prompt. Decisions should be grounded in objective, impactful distinctions, not subtle or
    debatable nuances.

* **Visual Fidelity:** Evaluate the technical quality and aesthetic alignment of the visuals,
    focusing on clarity, resolution (perceived vs. actual), unintended artifacts (e.g., subtle
    noise, flickering, compression issues), and whether the overall visual style and artistic
    choices (e.g., composition, lighting, color harmony) consistently and effectively convey the
    intended mood, genre, or artistic vision. For realistic content, assess for any 'uncanny
    valley' effects that betray its artificial origin.
* **Motions:** Evaluate the smoothness and naturalness of motion for all elements (e.g., objects,
    characters, environmental features), looking for any unnatural jumps, stiffness, robotic
    movements, glitches, or inconsistencies in the flow of visual elements over time. Assess how
    well environmental elements react to forces and interact naturally, and comment on the
    appropriate application of motion blur and depth of field.
* **Temporal Consistency:** Assess whether visual elements (e.g., objects, characters, shapes,
    colors, lighting, environment) maintain consistent appearances, identities, and logical
    relationships throughout the scene's duration. Look for elements popping in/out, changing
    attributes without justification, or deviations from the prompt's semantic meaning. Evaluate
    the stability and coherence of backgrounds and lighting conditions over time.
* **Audio Quality:** Evaluate the overall technical quality and aesthetic cohesion of all audio
    elements (e.g., dialogue, music, sound effects, ambience). Look for technical flaws (e.g.,
    hiss, clipping, distortion), and assess how well the sound elements are mixed, balanced, and
    contribute to the scene's intended mood and narrative. Consider clarity, richness, and
    artistic appropriateness of the soundscape, and whether audio elements are consistent in
    quality over time.
* **Audio-Video Alignment:** Assess how accurately audio events synchronize with corresponding
    visual actions and movements. Evaluate the effectiveness of audio spatializationÃ¢ÂĆ\ñâÃĬhow
    well sound conveys direction, distance, and the physical space of the scene. Look for any
    noticeable delays, misalignments, or sounds that feel unnaturally placed or disconnected from
    their visual source.
* **Prompt-Video Alignment:** Evaluate how accurately and completely the scene fulfills the
    specific content requirements of the scene prompt. Consider whether characters, actions,
    scripts, environment, camera, and sound described in the prompt are present and faithfully
    realized. Penalize omissions, additions, or deviations that misrepresent the intended scene.
* **Context Suitability:** Evaluate whether the character actions, environmental setting, events,
    and inferred dialogue are internally logical and plausible given their nature in the scene
    context. Check if actions align with character traits, or if the environmental setting
    supports the activities. Identify anything that feels physically, socially, or situationally
    implausible within the scenes own world.
* **Necessity:** Assess whether all elements in the scenes' characters, actions, objects,
    environmental details, and events are necessary and contribute meaningfully to the scene's
    core message.
* **Physical Commonsense:** Evaluate the physical presence of objects and actions in the scene
    that are unrealistic or break immersion, including anatomical errors (e.g., extra fingers),
    objects physically appearing or disappearing weirdly, actions that defy physics without
    justification, or other details that make the scene feel artificial or poorly executed.
* **Video Format:** Evaluate the visual resolution and smoothness of the first and last frames of
    the scene. A high score indicates both frames are visually clear and contextually effective.
* **Engagement:** Evaluate how effectively the scene captivates and retains viewer attention
    through compelling visuals, audio, and narrative elements. Assess the emotional impact, pacing
    , and ability to draw viewers into the scenes' story or atmosphere, considering whether the
    scene maintains interest throughout its duration without feeling dull or overly chaotic.
* **Safety:** Ensure the scene avoids visually harmful or inappropriate content, including graphic
    violence, sexually explicit imagery, self-harm depictions, disturbing visuals (e.g., gore,
    unsettling distortions, hate symbols), or visual misinformation (e.g., doctored images,
    misleading representations of real events). Flag any problematic visual elements.
* **Transition**: Evaluate the smoothness, coherence, and appropriateness of transitions between
    scenes, shots, or segments within the video or sequence. Assess whether transitions (e.g.,
    cuts, fades, dissolves, wipes) are abrupt, unpleasant, or visually and contextually suitable.
```

```
For each criterion, output the decision and a short explanation following the JSON format below:
```json
{{
    "visual_fidelity": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "motions": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "temporal_consistency": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "audio_quality": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "av_alignment": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "tv_alignment": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "context_suitability": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "necessity": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "scene_format": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "physical_commonsense": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "safety": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "engagement": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}},
    "transition": {{"Decision": "A_BETTER", "B_BETTER", or "TIE", "Explanation": ...}}
}}
```
```

## B.7. Prompt for Simple Pairwise Video Comparison

```
"""Which of the two videos more effectively addresses the following user prompt?
{input_prompt}

Please provide a brief explanation for your choice, and then indicate the final decision in the
    following format:

```json
{{
"Decision": "<A_BETTER | B_BETTER | COMPARABLE>"
}}```
"""
```

C. Additional Details

C.1. Human Evaluation Instructions

Self-Improvement Scoring Guidelines. Please evaluate the self-improooving trajectories below by assigning a score from **1 (Completely Worse)** to **5 (Completely Better)** according to the following guidelines:

- **1 – Completely Worse**: All self-improved videos are clearly worse than the initial video (4 out of 4 videos worsen).

- **2 – Marginally Worse**: The self-improved videos are generally worse than the initial video (at most 2 out of 4 videos worsen).

- **3 – Marginally Better**: Mixed results—some videos improve slightly, others worsen or remain the same. Overall a slight feeling of improvement.

- **4 – Better**: The self-improved videos are generally better than the initial video (at least 2 out of 4 videos show improvement).

- **5 – Completely Better**: All self-improved videos are clearly better than the initial video (4 out of 4 videos improve).

Visual Quality Scoring Guidelines. You will watch the following videos generated by an AI system. For each video, please evaluate its overall visual quality on a scale from **1 (Very Poor)** to **5 (Excellent)** according to the following guidelines:

- **5 – Excellent**: Very clear, sharp, natural, and pleasant to watch. No noticeable artifacts, distortions, or inconsistencies.

- **4 – Good**: Clear and understandable with only minor imperfections or faint artifacts.

- **3 – Fair**: Generally watchable, but noticeable artifacts, distortions, flickering, or inconsistencies reduce visual quality.

- **2 – Poor**: Difficult to watch, with significant visual artifacts, distortions, or temporal issues, though still recognizable.

- **1 – Very Poor**: Unwatchable or severely degraded. Strong artifacts, distortions, or glitches dominate the visuals.

Audio Quality Scoring Guidelines. You will listen to the following videos generated by an AI system. For each video, please evaluate its overall audio quality on a scale from **1 (Very Poor)** to **5 (Excellent)** according to the following guidelines:

- **5 – Excellent**: Very clear, natural, and pleasant to listen to. No noticeable noise, distortion, or artifacts.

- **4 – Good**: Clear and understandable with only minor imperfections or faint artifacts.

- **3 – Fair**: Generally understandable, but noticeable distortions, artifacts, or background issues reduce quality.

- **2 – Poor**: Difficult to listen to, with significant distortions, noise, or artifacts, though still recognizable.

- **1 – Very Poor**: Unintelligible or severely degraded. Strong artifacts, distortion, or noise dominate the audio.

D. Examples

D.1. Examples of VISTA's Optimized Prompts

Figures 12 and 13 present prompts optimized across iterations by VISTA.

D.2. VISTA's Behaviors

Through human evaluation, we identify key behaviors of VISTA that contribute to its significant improvements in video generation quality by refining prompts. First, **VISTA enhances video quality by improving prompt fidelity without inducing content drift**. As shown in Figure 14, compared to DP, the content of the generated videos remains faithful to the original prompt while achieving substantial quality gains. This improvement can be attributed to two key factors: the explicit constraints applied during the planning process, and the critique and prompt optimization mechanisms that enforce

Single-scene Dataset (Polyak et al., 2025): A spaceship entering hyperdrive, stars streaking past as it accelerates.

↓

VISTA at iteration 1: A sleek spaceship accelerates through space, its engines glowing brighter as stars begin to streak past. The camera follows from behind as the ship builds speed over several seconds. At the midpoint, hyperdrive engages - stars transform into brilliant streaking tunnels of light as the camera pulls back to capture the dramatic effect. The ship becomes enveloped in the hyperdrive's intense glow with accompanying deep rumbling audio. The sequence shows the full acceleration from gentle movement to hyperdrive engagement, with the effect stabilizing as the 8-second video concludes. Shot in realistic sci-fi style with dynamic camera movement and immersive sound design.

↓

VISTA at iteration 2: Create an 8-second, live-action, realistic sci-fi action video. The sequence initiates with a medium shot of a sleek, advanced spaceship, its engines softly glowing, set against a static starfield. This background should feature diverse star densities and a distant, subtle galaxy arm to deepen the cosmic perspective. As the ship commences acceleration, the camera smoothly tracks its motion from a slightly elevated, rearward perspective. This acceleration is a deliberate, gradual build-up over the initial seconds, allowing the surrounding stars to dramatically, vividly conveying escalating speed. At approximately the 4-5 second mark, the hyperdrive engages. The camera swiftly retreats and zooms out, maintaining the accelerating vessel in sharp focus while highlighting the intense, tunnel-like streaking of stars. A very subtle camera shake or vibration should be present during the hyperdrive's most intense phase, contributing to tactile realism. The hyperdrive audio must include a distinct low-frequency rumble and a rich, layered texture for maximum immersion. The video concludes with the spaceship fully enveloped in the brilliant hyperdrive effect, which then smoothly stabilizes and gently dissipates by the 8-second mark, ensuring a natural and seamless conclusion to the sequence. The video must be free of text overlays and human narration.

↓

VISTA at iteration 3:. Create an 8-second, live-action, realistic sci-fi action video. The sequence initiates with a medium shot of a sleek, advanced spaceship, its engines beginning with a subtle glow that gradually intensifies, set against a vast, static starfield. This cosmic backdrop features diverse star densities, a distant, subtle galaxy arm, and a very subtle parallax effect to enhance depth perception. An ambient space hum is faintly audible. As the ship commences acceleration, the camera smoothly tracks its motion from a slightly elevated, rearward perspective. This acceleration is a deliberate, gradual build-up over the initial seconds, allowing the surrounding stars to dramatically streak, vividly conveying escalating speed. Precisely at the 4.5-second mark, the hyperdrive engages. The camera swiftly retreats and zooms out, maintaining the accelerating vessel in sharp focus while highlighting the intense, tunnel-like streaking of stars. Dynamic lighting on the ship subtly reacts to the intense streaking. A perceptible but minimal camera shake or vibration is present during the hyperdrive's most intense phase, contributing to tactile realism. The hyperdrive audio includes a distinct low-frequency rumble and a rich, layered texture, with enhanced sonic complexity during the sustained phase for maximum immersion. The video concludes with the spaceship fully enveloped in the brilliant hyperdrive effect, which then begins its smooth stabilization and gentle dissipation immediately after its peak intensity, ensuring this process is clearly visible and progressing towards a natural and seamless conclusion by the 8-second mark. The video must be free of text overlays and human narration.

Single-scene Dataset (Polyak et al., 2025): The person's forehead creased with worry as he listened to bad news."

↓

VISTA at iteration 1: Produce an 8-second photorealistic video clip illustrating the user\\'s input: \"The person\\'s forehead creased with worry as he listened to bad news.\"\\n\\n\\nClip Details:\\n[0s\u20138s]:\\n   - Type: Character\\'s emotional response to receiving news.\\n   - Shot Progression: The scene opens with a medium shot of a man seated at a desk or table, holding a phone to his ear, his back slightly turned to the camera. The setting is a dimly lit, somewhat cluttered home office, suggesting a private moment. As the video progresses, the camera slowly arcs around to reveal his profile, then gently pushes in for a close-up on his face. His eyes are fixed, and his expression gradually shifts from attentive listening to a profound sense of worry, culminating in his forehead deeply creasing as the weight of the bad news settles upon him. The phone remains at his ear, a silent conduit of distress.\\n   - Tone: Anxious, somber, burdened.\\n   - Audio Details: The soundscape should be subtle and natural, primarily featuring ambient room tone, perhaps a very faint, almost imperceptible hum from the phone. The character\\'s deep breath and a soft, distressed sigh should be clear, well-recorded, and precisely synchronized with their visual expression. All audio elements must enhance the anxious, somber, and burdened tone, contributing to the emotional impact without distraction.\\n\\n\\nRestrictions:\\n- The video must be photorealistic and obey real-world physics, unless the User Prompt explicitly indicates otherwise.\\n- Do not include multiple scene changes or cuts, unless specified in the Clip Details.\\n- No captions or textual overlays.\\n- No spoken dialogue. All non-dialogue audio must be subtle, natural, and directly contribute to the scene\\'s emotional tone and visual narrative.

↓

VISTA at iteration 2: Produce an 8-second photorealistic video clip illustrating the user\\'s input: \"The person\\'s forehead creased with worry as he listened to bad news.\"\\n\\n\\nClip Details:\\n[0s\u20138s]:\\n - Type: Character\\'s emotional response to receiving news.\\n - Shot Progression: The scene opens with a medium shot of a man seated at a desk or table, holding a phone to his ear, his back slightly turned to the camera. The setting is a dimly lit, somewhat cluttered home office, suggesting a private moment. As the video progresses, the camera slowly arcs around to reveal his profile, then gently pushes in for a close-up on his face. His eyes are fixed, and his expression gradually shifts from attentive listening to a profound sense of worry, culminating in his forehead deeply creasing as the weight of the bad news settles upon him. The phone remains at his ear, a silent conduit of distress.\\n - Tone: Anxious, somber, burdened.\\n - Audio Details: The soundscape should be subtle and natural, primarily featuring the hushed ambient room tone of a quiet, enclosed home office, along with a very faint, low-frequency electronic hum emanating subtly from the phone. Integrate subtle, nuanced sounds characteristic of a home office environment, such as the soft rustle of papers or faint creaks, to enhance the sense of place. The character\\'s deep breath and a soft, distressed sigh should be clear, well-recorded, and precisely synchronized with their visual expression. Ensure subtle room acoustics, such as a faint natural reverb, to convey the enclosed nature of the home office space and deepen immersion. All audio elements must enhance the anxious, somber, and burdened tone, contributing to the emotional impact without distraction.\\n\\n\\nRestrictions:\\n- The video must be photorealistic and obey real-world physics, unless the User Prompt explicitly indicates otherwise.\\n- Do not include multiple scene changes or cuts, unless specified in the Clip Details.\\n- No captions or textual overlays.\\n- No spoken dialogue. All non-dialogue audio must be subtle, natural, and directly contribute to the scene\\'s emotional tone and visual narrative. Avoid any generic, abrupt, or uncontextualized sounds (e.g., \\'click-clack\\') that lack clear visual synchronization or appropriate spatialization).\\n- Ensure a smooth, gentle opening.\\n- The conclusion must be polished and complete, avoiding any abrupt cuts.

↓

VISTA at iteration 3: Generate an 8-second photorealistic video depicting a man's intense emotional reaction upon hearing distressing news. The video commences with a medium shot of the man in a dimly lit, slightly disheveled home office. He is seated at a desk, his back partially turned, a phone held firmly to his ear. A slow camera arc reveals his profile, then a gentle push-in brings the viewer to a close-up. His expression subtly shifts from focused attention to profound worry, culminating in a deeply furrowed brow as the gravity of the news registers. For the entire 8-second duration, the phone remains perfectly still and pressed against his ear, serving as a silent, unmoving conduit of distress, visually static without any shifts or adjustments. The overall tone should be anxious, somber, and heavy. The sound design should be subtle and natural, primarily featuring the hushed ambient sounds of the quiet, enclosed home office. A very faint, almost imperceptible, low-frequency electronic hum should consistently emanate from the phone. Subtle environmental sounds, such as soft paper rustles or faint creaks, should enhance the sense of place without becoming distracting. The character's deep, distressed breath and a soft sigh must be clear, well-recorded, and precisely synchronized with his visual reaction. Subtle room acoustics, including a faint natural reverb, should effectively convey the enclosed nature of the space. All audio elements must be understated, enhancing emotional impact without distraction. Crucially, no sounds related to phone handling or movement are permitted. The video must adhere to real-world physics, contain no scene changes or cuts, and be devoid of captions, textual overlays, or spoken dialogue. It requires a smooth, gentle opening and a polished, complete conclusion.

Figure 12 | Examples of prompts optimized by VISTA across iterations. Blue parts are updated.

Multi-scene Dataset: An 8-second video begins outdoors on a bright, sunny day. A bearded man in a red cap, blue-tinted sunglasses, purple hoodie, and black headphones addresses the camera in a static, chest-up shot. The suburban street background is softly blurred, showing houses and greenery under a clear blue sky. His voice is clear as he poses the question "Which comedian is known for their deadpan delivery?" A black text box with white lettering appears at the bottom displaying the question. After a pause, he states "Jeff Dye" with a knowing smile. The text smoothly transitions to show "JEFF 'DYE'". At 5.5 seconds, the video fades to a minimalist white outro featuring "Master of Puns" text with a 3D glasses emoji above a "SUBSCRIBE!" button. A crisp pop sound concludes the video.

VISTA at iteration 1: {'overall_content': 'The video features a man outdoors, asking a trivia question about a comedian known for deadpan delivery, and then immediately providing the answer. It concludes with a simple, branded outro screen encouraging viewers to subscribe.', 'theme': 'Trivia and lighthearted knowledge sharing, with a focus on comedy.', 'tone': 'Casual, informative, direct, and slightly humorous.' 'scenes': [{'timestamp': '0-5.5', 'duration_seconds': 5.5, 'scene_type': 'Man asking and answering a trivia question outdoors.', 'characters': 'A man with a full beard, wearing a red baseball cap, black sunglasses with blue reflective lenses, and a purple hoodie. He also has black over-ear headphones on.', 'actions': 'The man looks directly into the camera, speaking clearly. He first poses a trivia question, then pauses briefly before stating the answer. He offers a slight smile after delivering the answer.', 'dialogues': 'Man\'s voice: "Which comedian is known for their deadpan delivery?" (Text overlay appears: "WHICH COMEDIAN IS KNOWN FOR THEIR DEADPAN DELIVERY?") Man\'s voice: "Jeff Dye" (Text overlay changes to: "JEFF \"DYE\"")', 'visual_environment': 'A bright, sunny outdoor setting. The background is slightly blurred, showing residential houses with pitched roofs and some green foliage, suggesting a suburban street or sidewalk. The sky is clear and light blue. A prominent black rectangular text box with white text overlays the bottom portion of the screen.', 'camera': "A static, close-up shot of the man's face, framed from the chest up. The camera is held at eye level, giving a direct, personal feel.", 'sounds': 'Clear male voice speaking. Faint ambient street noise, possibly distant traffic or general outdoor sounds, is audible in the background.', 'moods': 'Informative, casual, engaging, and direct.'}

VISTA at iteration 2: An 8-second video unfolds, starting outdoors on a lively, sunlit day. For the first 5.5 seconds, the camera captures a man in a static, chest-up shot. He is adorned with a full beard, a red baseball cap, black sunglasses featuring striking blue reflective lenses, a purple hoodie, and black over-ear headphones. Behind him, a gently blurred backdrop showcases a serene suburban street with residential homes and abundant green foliage, all beneath a clear, light blue sky. The reflections visible in his sunglasses are flawlessly clean, ensuring no camera equipment or unwanted details are present to break the visual immersion. The man directly addresses the viewer, his voice clear and unaffected by any wind. A soft, consistent hum, suggesting distant street activity or general outdoor sounds, subtly enhances the natural ambiance without becoming a distraction. He asks, \"Which comedian is known for their deadpan delivery?\" Simultaneously, a sleek, black rectangular text box with white lettering gracefully appears from the bottom, positioned subtly at the screen's center-bottom, displaying \"WHICH COMEDIAN IS KNOWN FOR THEIR DEADPAN DELIVERY?\". After a short, expectant pause, he states, \"Jeff Dye.\" The text overlay then smoothly transitions: the question fades out as \"JEFF \\\"DYE\\\"\" fades in, maintaining its discreet size and placement. A subtle, knowing smile graces his lips as he concludes. At the 5.5-second mark, the video smoothly transitions from the live-action segment to the outro, utilizing a quick fade to black followed by a fade in. The final 2.5 seconds feature a pristine, minimalist white background. \"Master of Puns\" in black text, accompanied by a colorful 3D glasses emoji, is centered above a black rectangular \"SUBSCRIBE!\" button with a subtle grey border. A crisp 'pop' sound effect marks the video's conclusion."

VISTA at iteration 3: The 8-second video opens outdoors on a vibrant, sunlit day. For the initial 5.5 seconds, the camera presents a static, chest-up, eye-level view of a man. He sports a full beard, a red baseball cap, black sunglasses with striking blue reflective lenses, a purple hoodie, and black over-ear headphones. Behind him, a gently blurred scene of a serene suburban street, complete with residential homes and abundant green foliage, stretches beneath a clear, light blue sky. The reflections in his sunglasses are impeccably clean, free of any camera equipment or unwanted details, preserving visual immersion. The man addresses the viewer directly, his voice clear and unaffected by wind. A soft, consistent hum, hinting at distant street activity or general outdoor sounds, subtly enhances the natural ambiance without distraction. He asks, \"Which comedian is known for their deadpan delivery?\" Simultaneously, a sleek, black rectangular text box with white lettering gracefully emerges from the bottom, positioned subtly at the screen's center-bottom, displaying \"WHICH COMEDIAN IS KNOWN FOR THEIR DEADPAN DELIVERY?\". After a brief, expectant pause, he declares, \"Jeff Dye.\" The text overlay then smoothly shifts: the question fades out as \"JEFF \\\"DYE\\\"\" fades in, maintaining its discreet size and placement. A subtle, knowing smile plays on his lips as he finishes. At the 5.5-second mark, the video transitions from the live-action segment to the outro, marked by a quick fade to black followed by a fade in. The final 2.5 seconds showcase a pristine, minimalist white background. \"Master of Puns\" in black text, accompanied by a colorful 3D glasses emoji, is centered above a black rectangular \"SUBSCRIBE!\" button with a subtle grey border. A crisp 'pop' sound effect signals the video's conclusion.

VISTA at iteration 4: {'prompt_name': \"Trivia Master's Outdoor Riddle\", 'prompt_description': 'The 8-second video opens outdoors on a vibrant, sunlit day. For the initial 5.5 seconds, the camera presents a static, chest-up, eye-level view of a man. The composition is artfully balanced, utilizing the shallow depth of field to enhance the subject\\\'s presence and create a visually engaging frame. He sports a full beard, a red baseball cap, black sunglasses with striking blue reflective lenses, a purple hoodie, and black over-ear headphones. Behind him, a gently blurred scene of a serene suburban street with residential homes and abundant green foliage, stretches beneath a clear, light blue sky. The bright, sunny outdoor setting is bathed in a warm, inviting light, creating a subtly cinematic atmosphere. The reflections in his sunglasses are impeccably clean, free of any camera equipment or unwanted details, preserving visual immersion.\\\n\\\nA clear, warm, and resonant male voice speaks directly to the viewer. Faint, yet subtly varied, ambient street noise, including distant traffic and occasional natural outdoor sounds (e.g., a faint bird chirp or rustle of leaves), is audible in the background, adding to the immersive realism. The man asks, \"Which comedian is known for their deadpan delivery?\" Simultaneously, a sleek, black rectangular text box with white lettering gracefully emerges from the bottom, positioned subtly at the screen\\\'s center-bottom, displaying \"WHICH COMEDIAN IS KNOWN FOR THEIR DEADPAN DELIVERY?\". After a brief, expectant pause, he declares, \"Jeff Dye.\" The text overlay then smoothly shifts: the question fades out as \"JEFF \\\\\\"DYE\\\\\\"\" fades in, maintaining its discreet size and placement. A subtle, knowing smile plays on his lips as he finishes.\\\n\\\nAt the 5.5-second mark, the video transitions from the live-action segment to the outro, marked by a quick fade to black followed by a fade in. The final 2.5 seconds showcase a pristine, minimalist white background. \"Master of Puns\" in black text, accompanied by a colorful 3D glasses emoji, is centered above a black rectangular \"SUBSCRIBE!\" button with a subtle grey border. A crisp \\\'pop\\\' sound effect signals the video\\\'s conclusion.'}"

Figure 13 | Examples of prompts optimized by VISTA across iterations. Blue parts are updated.

the text-video alignment. Second, **VISTA significantly improves instruction-following in SOTA video generation models**. As seen in Figure 14, DP often fails to meet prompt specifications, while VISTA successfully corrects such failures. This improvement stems from VISTA's strict enforcement of text-video alignment during video selection and its use of feedback that evaluates alignment and contextual relevance. Finally, **VISTA reduces physical, visual, and audio hallucinations**. While models like Veo 3 often produce videos with abrupt object changes, implausible motions, and unsolicited audio or text, VISTA mitigates these issues through constraint-guided selection and strict penalties for violations (Alg. 1-Step 2).

DP generates video with gremlins moving backward fast without wooden rollercoaster, which is physically non-sense.

VISTA fixes the issues with better visual fidelity: gremlines are moving forward and camera is backward.

Prompt: A rapid tracking shot of small, big-eared **gremlins on a wooden rollercoaster** in a midcentury theme park...

DP fails to cut between the character's face shot and the interior ceiling shot, nor repeat this transition.

VISTA fixes this with an even better camera focus and more natural character's expression.

Prompt: A short, humorous video depicting a young woman's mood rapidly shifting from bored and slightly annoyed to overtly joyful, triggered by a sudden and dramatic change in background music...
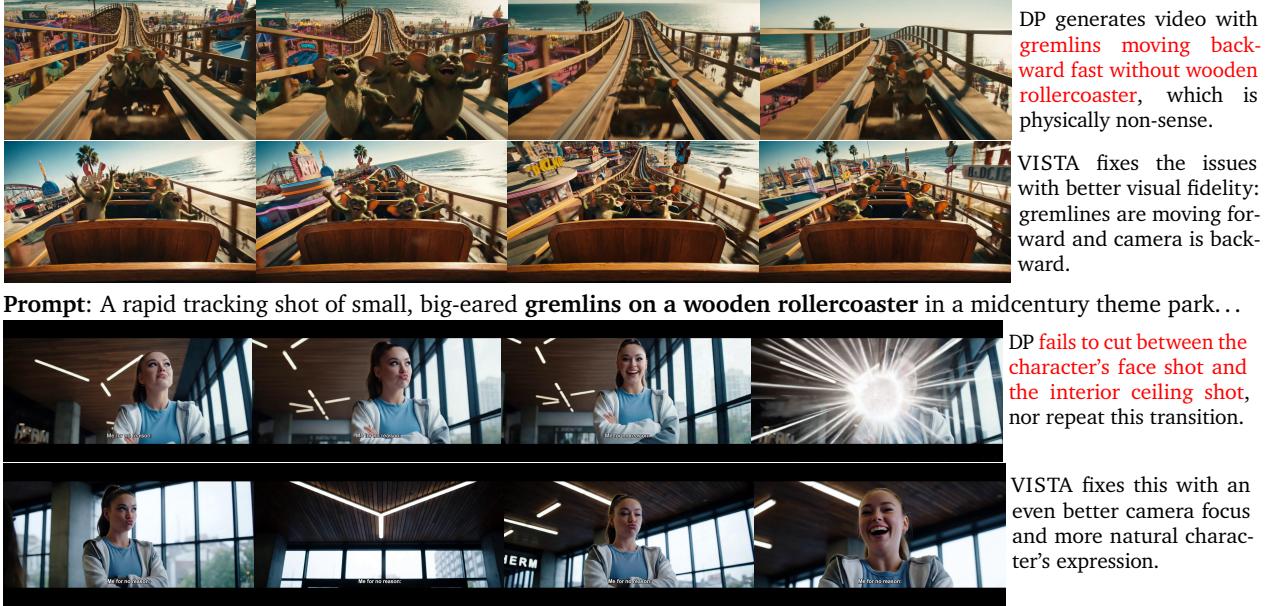
Figure 14 | Comparison of Direct Prompting (DP) and VISTA with Veo 3. Top rows show Veo 3 failing to follow human instructions and generating unrealistic gremlin motions. Bottom rows highlight its inability to produce coherent multi-scene videos under DP. In contrast, VISTA corrects these issues with more realistic motions and smoother visual transitions.

D.3. Video Examples

We invite audiences to visit our project page and Table 1 for more video examples.

D.4. MMAC Examples

Section D.4 shows more examples of critiques from VISTA compared to Visual Self-Refine.

D.5. Deep-Thinking Prompting Agent

Recall the video from Table 9:

Deep-Thinking Procedure Answers:
1. Review the Issues
Upon a thorough review of the provided feedback, several areas for improvement have been identified, indicated by qualitative feedback suggesting less-than-optimal performance, even if a numerical score isn't explicitly given as "less than 8." The prompt asks to identify issues with scores less than 8, but since numerical scores aren't provided, I will interpret "major issues" as those with detailed negative qualitative feedback.
The primary issues are:
* **Visual Fidelity:** The feedback highlights "visible reflections of the camera setup in the subject's sunglasses." This breaks immersion and indicates a lack of control over subtle visual details. Furthermore, the "prominent size" of the text overlays and their "abrupt appearance" are noted as detracting from a smooth visual experience. While legible, their visual impact is not ideal. * **Motions and Dynamics:** This category specifically calls out the "instantaneous appearance and disappearance of the text overlays" as abrupt, reiterating the point from visual fidelity. This indicates a lack of fluid animation for on-screen graphics, which impacts the overall polish and dynamic flow of the video. * **Audio Quality Cohesion:** A significant concern is the "noticeable wind noise throughout the outdoor segment,"

Prompt: A spaceship entering hyperdrive, stars streaking past as it accelerates.

Self-Refine: The video is a strong visual representation of the prompt. The criticisms are minor and relate more to stylistic choices or common sci-fi tropes rather than a failure to meet the core request. It successfully conveys a spaceship accelerating into a hyperdrive state with stars streaking past.

VISTA (Visual Fidelity): The Normal Judge praises the high technical quality, clarity, sharp details, and consistent sci-fi aesthetic, noting the well-executed lighting and vibrant engine glow. They suggest minor improvements like more dynamic lighting and nuanced atmospheric distortion. The Negative Judge, however, critically points out uninspired lighting on the spaceship, and static background galaxy, and overly uniform, artificial-looking streaking stars lacking variation or parallax. They also mention the abruptness of transitions affecting fidelity. Upon review, the Negative Judge's assessment carries more weight. While the video is technically clean, the lighting on the ship does appear quite uniform, lacking the dynamic interplay of light and shadow expected in space...

Self-Refine: Overall, the generated video is **highly successful** in fulfilling the prompt. It accurately and visually compellingly depicts a spaceship entering hyperdrive with stars streaking past as it accelerates. There are no significant "wrongs" or misinterpretations of the prompt. The minor points above are more about stylistic choices or slight enhancements rather than fundamental errors.

VISTA (Motions and Dynamics): While the Normal Judge praises the smoothness of the ship's motion and dynamic star streaking, the Negative Judge correctly identifies a major directional flaw: the spaceship moves vertically, which conflicts with viewer expectations of horizontal acceleration. Additionally, the Negative Judge points out the lack of micro-dynamics (e.g., rotational drift, buildup phases) and unrealistic exhaust behavior, which diminish the believability of motion. These omissions outweigh the surface-level smoothness.

Prompt: The couple runs hand in hand to release a sky lantern, then watches it drift upward into the night sky, carried by the wind with the stars shining above.

Self-Refine: The video successfully portrays the actions of the couple running and releasing a lantern. However, it **fails significantly on the atmospheric detail of the "night sky with stars shining above."** The absence of stars is a major oversight that detracts from the prompt's intended mood and visual.

VISTA (Contextual Suitability): The scene struggles significantly with internal logical consistency, particularly regarding its temporal and environmental progression. An abrupt and jarring shift from a bright, late-afternoon setting to a deep, artificial blue night sky without any visual cues for time passing severely undermines immersion. This featureless 'night sky' further detracts from plausibility, failing to represent a natural environment. While individual actions are suitable, their placement within such an inconsistent context creates a sense of unreality.

VISTA (Visual Characters): While the two characters are central and necessary to the narrative, their portrayal lacks emotional depth and genuine interaction. Their expressions, particularly during the lantern release, appear somewhat static and posed, missing an opportunity to convey a deeper sense of wonder, intimacy, or shared experience. This underdevelopment makes them feel more like subjects performing actions rather than individuals experiencing a meaningful moment, limiting the scene's emotional resonance.

Figure 15 | VISTA delivers high-level feedback on visual fidelity, contextual consistency, and emotional expression, surfacing nuanced flaws (e.g., static lighting, implausible scene transitions, and lack of character depth) that conventional LLM critiques often miss but are readily noticed by human judges.

which compromises the professionalism and clarity of the audio. Additionally, the feedback points out a "lack of depth due to the absence of a subtle, natural ambient bed," making the scene feel "somewhat sterile" despite clear dialogue. This suggests the ambient sound generation was either insufficient or poorly integrated. * **Timeline and Transition:**

While the cut to the outro screen is described as "clean and intentional," the feedback explicitly states it is "not a 'fluid' transition (e.g., a dissolve or fade)." It suggests that "for absolute perfection in 'smoothness,' a subtle transition effect could have been employed." Although deemed "common and acceptable" for short-form content, it's highlighted as an area where "smoothness" could be improved, indicating it's not achieving the highest possible quality in this specific aspect.

These issues collectively point to a need for more granular control over visual effects, on-screen text animation, environmental audio generation, and scene transitions to elevate the video's overall production quality and polish. Addressing these will enhance viewer immersion and the professional feel of the content.

2. **Define the Objectives**

The overarching objective of this video, as defined by the User Prompt, is to create a short-form piece of content centered around "Trivia and lighthearted knowledge sharing, with a focus on comedy." The video aims to be "Casual, informative, direct, and slightly humorous" in tone. The specific structure involves a man asking and answering a trivia question outdoors, followed by a branded outro screen with a call to action. The video has a strict length constraint of 8 seconds, with a precise breakdown of scene durations (5.5 seconds for the trivia segment and 2.5 seconds for the outro).

The prompt implicitly sets high expectations for visual and audio fidelity, given the detailed descriptions of character appearance, actions, visual environment, camera work, and sound design. Success criteria include clear dialogue, appropriate visual framing, a natural outdoor setting, and a clean, effective call-to-action screen. The target audience appears to be general viewers interested in quick, engaging trivia content, likely on social media platforms where short, polished videos perform well. The key message is to share a piece of trivia and encourage viewer engagement through subscription. The prompt's specificity across various elements (e.g., "red baseball cap," "black sunglasses with blue reflective lenses," "purple hoodie," "black rectangular text box with white text") indicates a desire for precise execution and a high degree of alignment with the provided specifications. The feedback further reinforces this by praising "exceptional alignment" and "excellent temporal consistency," suggesting that the core objective of faithful reproduction was largely met, but refinement in specific areas is still needed for optimal quality.

3. **Identify Model Limitations, Given the Video Prompt**

Analyzing the identified major issues in conjunction with the Video Prompt, several points suggest potential limitations of the underlying video generation model when interpreting less explicit instructions or handling complex real-world phenomena.

* **Reflections in Sunglasses:** The presence of "visible reflections of the camera setup in the subject's sunglasses" points to a challenge in rendering complex optical phenomena accurately without explicit guidance. Simulating realistic reflections, especially on curved, reflective surfaces like sunglasses, requires sophisticated ray tracing or similar techniques. Without a specific instruction to "minimize reflections" or "ensure no camera reflections are visible," the model might default to a physically plausible but undesirable outcome. This is a common limitation in generative models that aim for realism but may not prioritize specific aesthetic refinements unless prompted. * **Abrupt Text Overlays (Appearance/Disappearance Size):** The "instantaneous appearance and disappearance" of text overlays and their "prominent size" suggest the model's default behavior for on-screen graphics. When the prompt simply states "Text overlay appears" or "Text overlay changes," the model interprets this as a binary state change (on/off) rather than a smooth animation (e.g., fade, slide). Similarly, without explicit size or placement parameters, the model might choose a default that is visually impactful but not aesthetically subtle. This indicates a limitation in the model's ability to infer desired animation styles or optimal visual hierarchy for text without explicit instructions. * **Wind Noise and Lack of Ambient Depth:** The "noticeable wind noise" and absence of a "subtle, natural ambient bed" highlight a limitation in generating nuanced and clean environmental audio. While the prompt asks for "Faint ambient street noise," generating specific, clean ambient sounds while simultaneously filtering out undesirable elements like wind noise (which is often present in outdoor recordings) is a complex audio engineering task. The model might struggle to differentiate between desired "outdoor sounds" and distracting "noise," or to layer sounds effectively to create depth without explicit instructions on sound mixing, noise reduction, or specific ambient sound profiles. * **Non-Fluid Transition:** The "immediate cut" from the live-action scene to the outro, despite the feedback suggesting a "subtle transition effect" for "absolute perfection," indicates that the model defaults to a hard cut when no specific transition type is mentioned. This is a common behavior for generative models; they typically require explicit instructions for dissolves, fades, or other cinematic transitions, as a hard cut is the simplest and most direct way to move between scenes.

These issues are not necessarily "failures" of the model but rather areas where its default interpretations or capabilities fall short of achieving a highly polished, professional output without more precise and detailed prompting.

4. **Identify Video Prompt Issues, Given the Model**

Considering the identified model limitations and the feedback, the Video Prompt exhibits several areas where its vagueness, lack of specificity, or missing information contribute directly to the observed issues.

* **Vague Text Overlay Instructions:** The prompt states "Text overlay appears" and "Text overlay changes." This is a vague instruction regarding the *manner* of appearance. It doesn't specify any animation (e.g., fade-in, slide-up, dissolve). This vagueness leads the model to default to an abrupt, instantaneous appearance and disappearance, as noted in the feedback for both "visual fidelity" and "motions and dynamics." Furthermore, the prompt does not specify the *size* or *prominence* of the text overlays, leading to them occupying a "significant portion of the screen," which

was deemed too prominent. * **Insufficient Audio Detail:** The instruction "Faint ambient street noise, possibly distant traffic or general outdoor sounds, is audible in the background" is too broad. It doesn't explicitly request noise reduction (e.g., for wind) or specify the *quality* or *composition* of the ambient soundscape (e.g., "subtle, consistent ambient bed"). This lack of detail allows the model to generate "noticeable wind noise" and a sterile soundscape lacking depth, as it interprets "outdoor sounds" without the necessary refinement or filtering instructions. * **Missing Transition Specification:** The prompt defines the timestamps for the two scenes but provides no instruction whatsoever regarding the *type* of transition between the live-action segment and the outro screen. This missing information leads the model to implement a default hard cut, which, while functional, is not considered "fluid" or "perfectly smooth" according to the feedback. To achieve a more polished transition, a specific instruction (e.g., "subtle cross-dissolve") is required. * **Lack of Aesthetic Refinement Instructions:** The prompt does not include any instructions related to minimizing undesirable visual artifacts like reflections in eyewear. While it describes the character's attire, it doesn't anticipate or address potential issues with reflective surfaces. This missing aesthetic constraint allows the model to generate "visible reflections of the camera setup," which detract from immersion.

In essence, the prompt, while highly detailed in many aspects, lacks the necessary specificity in areas related to visual animation, audio refinement, and subtle aesthetic controls. These omissions force the model to make assumptions or default to basic implementations that do not meet the desired level of polish and fluidity.

5. **Propose Targeted Revisions to the Video Prompt**

Based on the comprehensive analysis of the issues, model limitations, and prompt shortcomings, the following targeted revisions are proposed to enhance the Video Prompt's effectiveness and guide the model toward generating a higher-quality, more polished video:

1. **For Visual Fidelity (Reflections in Sunglasses):** * **Modification Action:** Add a specific instruction within the character description or visual environment section to minimize or avoid reflections of the camera setup in the subject's sunglasses. This could be phrased as a constraint or a desired aesthetic outcome. * **Rationale:** This directly addresses the feedback regarding "visible reflections of the camera setup," which breaks immersion. By explicitly stating this, the prompt guides the model to prioritize a cleaner visual on reflective surfaces.

2. **For Visual Fidelity & Motions/Dynamics (Text Overlays):** * **Modification Action:** Modify the dialogue section for the text overlays to specify a smooth animation for their appearance and disappearance (e.g., "fade in" or "slide up from bottom"). Also, add a constraint on their size and placement to ensure they are legible but not overly prominent (e.g., "subtle, occupying minimal screen space at the bottom center"). * **Rationale:** This addresses the "abrupt appearance" and "instantaneous disappearance" of text, as well as their "prominent size." Specifying animation will create a smoother visual flow, and defining size/placement will ensure they serve their purpose without being distracting.

3. **For Audio Quality Cohesion (Wind Noise & Ambient Sound):** * **Modification Action:** Refine the "sounds" description for the first scene to explicitly request that dialogue be free of wind noise and to specify a subtle, consistent ambient sound bed. For example, "Clear male voice speaking, free of wind noise. A subtle, consistent ambient street soundscape, such as distant traffic or general outdoor sounds, is audible in the background, ensuring depth without distraction." * **Rationale:** This directly tackles the "noticeable wind noise" and the "lack of depth" in the audio. By being more prescriptive about noise reduction and the quality of ambient sound, the prompt guides the model to generate a cleaner and richer audio experience.

4. **For Timeline and Transition (Scene Transition):** * **Modification Action:** Add a specific instruction for the transition between the first scene (0-5.5s) and the second scene (5.5-8s). Specify a subtle, fluid transition type, such as a "subtle cross-dissolve" or "quick fade to black then fade in." * **Rationale:** This addresses the feedback that the immediate cut, while acceptable, is not "fluid." By explicitly requesting a subtle transition, the prompt ensures a more polished and seamless flow between the live-action content and the call-to-action screen, enhancing the overall production quality.

These revisions aim to provide the model with clearer, more specific instructions in areas where its default interpretations fell short of the desired polished outcome, without overly complicating the core content or structure of the video.

6. **Revise the Suggested Modification Actions**

Upon reviewing the proposed modifications against all the major issues identified, the suggested actions comprehensively address each point. The revisions are targeted, specific, and aim to provide the model with clearer guidance where the original prompt was vague or lacked detail.

* **Reflections in sunglasses:** The action to "Add a specific instruction within the character description or visual environment section to minimize or avoid reflections of the camera setup in the subject's sunglasses" directly addresses this visual fidelity issue. * **Abrupt text overlays and prominent size:** The action to "Modify the dialogue section for the text overlays to specify a smooth animation for their appearance and disappearance (e.g., 'fade in' or 'slide up from bottom'). Also, add a constraint on their size and placement to ensure they are legible but not overly prominent (e.g., 'subtle, occupying minimal screen space at the bottom center')" directly addresses both the motion/dynamics and visual fidelity concerns related to text. * **Wind noise and lack of ambient depth:** The action to "Refine the 'sounds' description for the first scene to explicitly request that dialogue be free of wind noise and to specify a subtle, consistent ambient sound bed. For example, 'Clear male voice speaking, free of wind noise. A subtle, consistent ambient street soundscape, such as distant traffic or general outdoor sounds, is audible in the background, ensuring depth and

naturalism without distraction."' directly addresses the audio quality cohesion issues. * **Non-fluid transition:** The action to "Add a specific instruction for the transition between the first scene (0-5.5s) and the second scene (5.5-8s). Specify a subtle, fluid transition type, such as a 'subtle cross-dissolve' or 'quick fade to black then fade in.'" directly addresses the timeline and transition feedback.

All major issues are covered. The modifications are actionable and aim to improve the video's polish and professional quality without introducing unnecessary complexity or deviating from the original intent. No further revisions to the suggested modifications are needed as they are precise and comprehensive.

```python

["Modify the 'characters' or 'visual_environment' description for the first scene to include a constraint: 'Ensure reflections in the subject's sunglasses do not show camera equipment or distracting elements, maintaining visual immersion.'", "Update the 'dialogues' section for the first scene's text overlays: 'Text overlay appears: WHICH COMEDIAN IS KNOWN FOR THEIR DEADPAN DELIVERY?'(Text overlay should smoothly fade in/slide up from the bottom, be legible, and occupy a subtle, non-prominent portion of the screen, ideally centered at the bottom.)' and 'Text overlay changes to: JEFF DYE (Text overlay should smoothly transition or fade out the previous text and fade in/slide up the new text, maintaining the subtle size and placement.)'", "Refine the 'sounds' description for the first scene: 'Clear male voice speaking, with dialogue free of noticeable wind noise. A subtle, consistent ambient street soundscape, such as distant traffic or general outdoor sounds, is audible in the background, ensuring depth and naturalism without distraction.'", "Add a specific instruction for the transition between the first scene (timestamp '0-5.5') and the second scene (timestamp '5.5-8'): 'The transition from the live-action scene to the outro screen should be a subtle cross-dissolve or a quick, smooth fade to black then fade in, rather than an immediate hard cut.'"]```