AFFECTVBENCH: BENCHMARKING AFFECTIVE FAITHFULNESS IN LONG-FORM TEXT-TO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

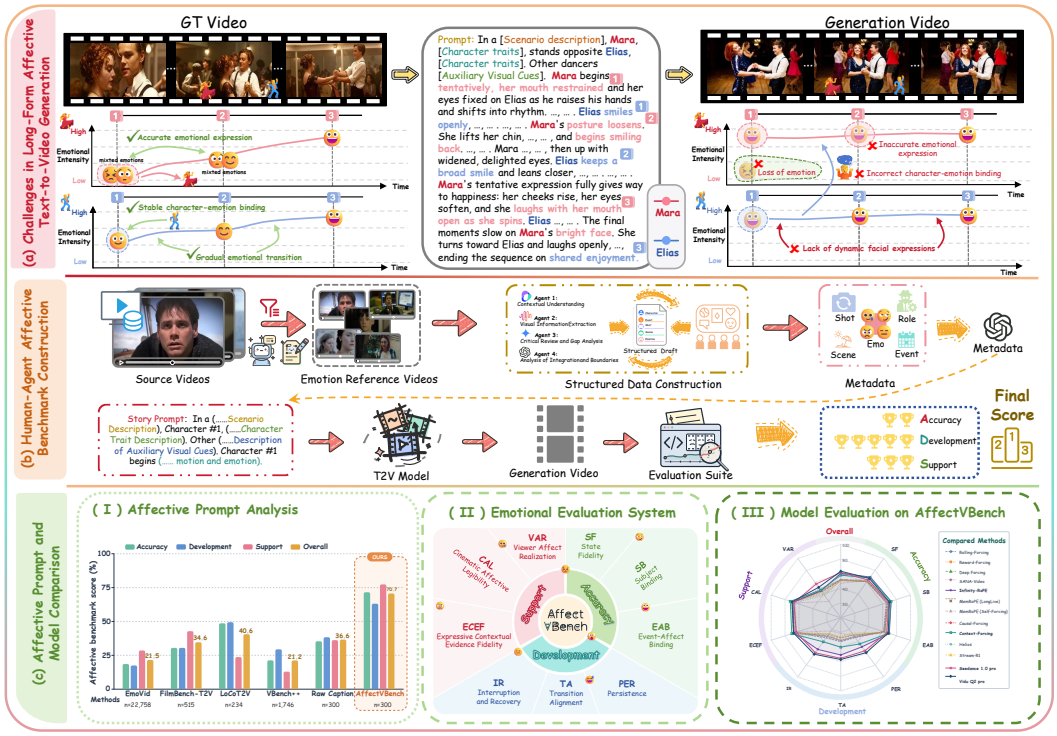


Figure 1: **Overview of AffectVBench.** (a) The pronounced affective gap between GT and generated videos motivates our AffectVBench. (b) A human–agent pipeline constructs an affect-rich benchmark from real videos with verified Story Prompts and structured metadata. (c) AffectVBench provides richer affective prompt coverage and enables fine-grained assessment of affective faithfulness.

ABSTRACT

Existing benchmarks have accelerated long-form text-to-video (T2V) generation by assessing visual quality, temporal coherence, and cinematic presentation, but overlook *affective faithfulness*: whether the intended affect is correctly bound to characters and events, develops coherently across shots, and is supported by perceivable evidence. We introduce **AffectVBench**, a fine-grained benchmark for this capability. Grounded in Story Prompts and structured references capturing open-vocabulary affective states, intensity, mixtures, and temporal evolution, AffectVBench provides nine evidence-grounded metrics across Accuracy, Development, and Support. It further incorporates blinded anchor-based calibration for Viewer Affect and multidimensional uncertainty-aware probabilistic scoring. Experiments on 13 representative T2V models reveal that current models struggle with multi-character emotion binding, cross-shot affect persistence, and event-driven affect transitions. These findings establish affective faithfulness as a distinct bottleneck for credible long-form video generation.

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

1 INTRODUCTION

Text-to-video (T2V) generation is advancing at an unprecedented pace, evolving from short, single-shot clips toward long-form narratives with multiple characters, events, and shots. This progress has been propelled not only by rapid technical innovation, but also by benchmarks that continually expose model limitations and steer subsequent development. Benchmarks such as VBench (Huang et al., 2024), LoCoT2V-Bench (Zheng et al., 2025), and FilmBench (Wang et al., 2026) have expanded evaluation from visual quality and text alignment to long-range coherence, narrative structure, and cinematic expression. Yet a defining gap remains between current T2V outputs and compelling professional cinema (e.g., award-winning cinema): the richness, coherence, and emotional resonance of affective expression. As illustrated in Figure 1 (a), a generated video may preserve recognizable characters and smooth shot transitions while expressing an incorrect emotion, assigning the intended emotion to the wrong character, losing an established emotional state across shots, or failing to visually realize an affective transition. Closing this gap first requires a principled measure of *affective faithfulness*, a critical yet largely overlooked problem in T2V evaluation.

Existing affect-related evaluations remain inadequate for this purpose. T2V benchmarks typically reduce affect to a global label, an isolated facial expression, or a component of perceptual quality, whereas affective video understanding methods recognize emotions already present in real videos rather than assess whether a specified affective narrative has been faithfully generated. *A rigorous evaluation must therefore address three challenges.* First, narrative affect is structurally complex: it is open-vocabulary, intensity-sensitive, potentially mixed, and dynamically shaped by unfolding events. Second, affective failures are heterogeneous: a model may realize the wrong state, bind it to the wrong character or event, disrupt its temporal development, or provide insufficient visual evidence, making a single aggregate score inadequate. Third, character affect and viewer affect are not equivalent. The latter is subjective and context-dependent, requiring calibrated references and uncertainty-aware scoring.

To address these challenges, we introduce **AffectVBench**, a fine-grained benchmark for affective faithfulness in long-form T2V generation. To capture open-ended and evolving affect, AffectVBench is grounded in real-world videos rather than a predefined emotion taxonomy. As summarized in Figure 1 (b), a multi-agent pipeline extracts characters, shots, narrative events, affective states, and their temporal dependencies, which are then verified and consolidated by humans into structured metadata and Story Prompts. These annotations capture affective states and intensity, mixed emotions, character and event bindings, and temporal evolution. To diagnose heterogeneous failures, AffectVBench organizes nine metrics into three dimensions, as shown in Figure 1 (c). ① **Accuracy** assesses what affect is realized and how it is bound (State Fidelity, Subject Binding, and Event–Affect Binding). ② **Development** evaluates how affect persists and evolves (Persistence, Transition Alignment, and Interruption and Recovery). ③ **Support** measures whether sufficient visual evidence makes the intended affect perceptible (Expressive Contextual Evidence Fidelity, Cinematic–Affective Legibility, and Viewer Affect Realization). Finally, to accommodate the subjectivity of Viewer Affect, we introduce blinded video-anchor calibration and uncertainty-aware probabilistic scoring, establishing a shared perceptual scale while retaining disagreement, insufficient evidence, and unjudgeable cases.

We evaluate 13 representative T2V generation methods using AffectVBench. Although current models produce visually coherent videos and locally plausible affective cues, they often fail to organize these cues into character-consistent, event-grounded, and temporally coherent trajectories, identifying affective faithfulness as a key bottleneck in long-form T2V generation.

Our contributions are threefold.

- We formulate *affective faithfulness* as a distinct evaluation problem to steering long-form T2V development and introduce AffectVBench, a real-video-grounded benchmark constructed through a human–agent pipeline.
- We develop nine fine-grained metrics across Accuracy, Development, and Support, together with blinded video-anchor calibration and uncertainty-aware scoring for Viewer Affect.
- We systematically evaluate 13 representative T2V methods, establishing a comprehensive baseline and revealing their limitations in realizing coherent affective trajectories.

2 Related Work

Long-form and Complex Text-to-Video Generation and Evaluation. T2V generation has advanced from short, single-shot clips to long-form narratives involving multiple shots, events, and characters. Current methods broadly follow two directions. Multi-prompt methods decompose narratives into prompt-conditioned segments and enforce cross-segment coherence (Teng et al., 2025; Yesiltepe et al., 2026; Li et al., 2026). Continuous-generation methods extend temporal context via autoregressive rollout, stateful memory, compression, or distribution correction (Liu et al., 2025; Chen et al., 2025; Zhu et al., 2026; Yuan et al., 2026). Concurrently, video evaluation benchmarks have accordingly expanded beyond perceptual quality. VBench (Huang et al., 2024) and VBench++ (Huang et al., 2025) evaluate visual quality, motion dynamics and text–video alignment, while T2V-CompBench (Sun et al., 2025) examines compositional constraints. For extended narratives, recent efforts such as MovieGenBench (Polyak et al., 2024), FilmBench (Wang et al., 2026), and LoCoT2V-Bench (Zheng et al., 2025) assess cinematic storytelling, cross-scene consistency, event completion, and temporal ordering. Despite these advances, existing benchmarks systematically overlook dynamic emotion, reducing it to a static global attribute or an isolated clip-level label, or omitting it entirely. In complex narrative generation, however, emotion is intrinsically character-bound and event-driven: while character identity demands cross-shot invariance, emotion must continuously evolve and transition along narrative trajectories, yet such dynamics remain untouched by standard consistency or alignment metrics. *To bridge this gap, we model emotion as a temporally traceable state conditioned on characters and narrative events, providing a systematic evaluation of emotional fidelity, developmental trajectories, and causal grounding.*

Emotion Representation and Affective Video Understanding. Affective video understanding characterizes expressed and evoked emotions using discrete categories (Jiang et al., 2020) or continuous dimensions (Baveye et al., 2015). Early multimodal benchmarks, including MELD (Poria et al., 2019), CMU-MOSEI (Zadeh et al., 2018), and CMU-MOSEAS (Zadeh et al., 2020), evaluate affective cues across visual, acoustic, and linguistic modalities. Recent multimodal large language models extend this scope to open-vocabulary semantics, causal reasoning, and dual affective perspectives. MME-Emotion (Zhang et al., 2026) benchmarks emotion recognition alongside trigger-oriented reasoning. EEmo-Bench (Gao et al., 2025a) evaluates image-evoked emotion and valence-arousal-dominance (VAD) attributes. E^3mo-Bench (Gao et al., 2026) covers expressed and evoked emotion through perception, open-vocabulary recognition, and VAD assessment. Despite their success in real-world scenarios, existing affective computing paradigms do not directly transfer to T2V evaluation. They assume natural facial motion and high-quality footage, making them sensitive to synthetic artifacts, temporal inconsistencies, and unnatural facial deformations. Moreover, their clip-level classification and post-hoc reasoning do not assess text–emotion alignment, character-specific binding, or narrative-driven emotional transitions. Conversely, general T2V benchmarks emphasize visual and motion quality while largely overlooking affect. *We therefore model emotion as a temporally traceable state grounded in narrative events and observable cues, shifting evaluation from isolated emotion recognition to text-grounded affective fidelity and dynamic emotional evolution.*

3 AffectVBench Construction

AffectVBench rigorously evaluates affective faithfulness in long-form video generation. We first describe the curation of source videos with diverse and discernible affective content. We then introduce a multimodal affective narrative construction pipeline that leverages multiple MLLMs (Yang et al., 2025; Wang et al., 2025) to identify and cross-validate characters, narrative events, temporal dependencies, and affective transitions from both audiovisual and visual-only streams. The resulting evidence is consolidated through human verification into structured narrative metadata that explicitly captures how characters, events, and affective states evolve and interact over time. Finally, this metadata is transformed into Story Prompts and corresponding evaluation references for systematic evaluation (Figure 1 (b)).

3.1 Source Video Curation

To capture the diversity, richness, and complexity of affective expression in long-form narratives, we curate character centered videos from MovieChat 1K (Song et al., 2024) and VEATIC (Ren et al.,

2024), prioritizing samples with diverse interactions, coherent event progression, and nontrivial affective dynamics. We use FFmpeg for video decoding, shot segmentation, temporal indexing, and representative frame extraction, followed by an initial screening based on character identifiability, event completeness, affective development, and the availability of observable visual evidence. Two reviewers then independently examine the major characters, key events, affective changes, and their visual grounding, removing samples with ambiguous character or event relations, insufficient affective development, or excessive dependence on dialogue and subtitles. Disagreements are resolved by a third reviewer. This process yields 300 candidate videos for subsequent Story Prompt construction, with an average duration of approximately 59 s, 10 shots, 2.3 major characters, and 5.2 narrative events per video.

3.2 Multimodal Affective Narrative Construction

Metadata Construction. To recover semantically complete yet visually grounded affective narratives, we construct complementary representations from the source videos. Seed2.1 Pro (Gao et al., 2025b) first analyzes the full audiovisual context, including visual content, dialogue, and subtitles, to recover characters, events, relationships, and affective triggers. In parallel, Qwen3.8 Max (Yang et al., 2025) independently examines the muted visual stream to identify visually observable characters, actions, events, and affective evidence. MiniMax M3 (Lai et al., 2026) then acts as a grounding critic to identify conflicts in character identity, temporal relations, affect attribution, and evidence support. GPT 5.6 sol (Singh et al., 2025) resolves these complementary analyses into structured narrative metadata while preserving the distinction between contextual semantics and visually grounded evidence. Human reviewers further verify character identities, key events, affect attribution, temporal evolution, and their supporting visual cues.

Story Prompt Generation. Based on the verified metadata, GPT 5.6 sol serves as a **Visual Story Compiler** that transforms structured narrative information into a coherent Story Prompt for silent T2V generation. The compilation preserves the original narrative while explicitly organizing character identities, event progression, affective states, temporal evolution, and observable expressive cues into visually realizable descriptions. Seed2.1 Turbo (Gao et al., 2025b) independently verifies source faithfulness, visual realizability, character binding, affective continuity, and dependence on nonvisual information. Human reviewers finally examine narrative completeness, overall coherence, and the natural progression of character affect to obtain the final Story Prompt.

4 AffectVBench Evaluation Suite

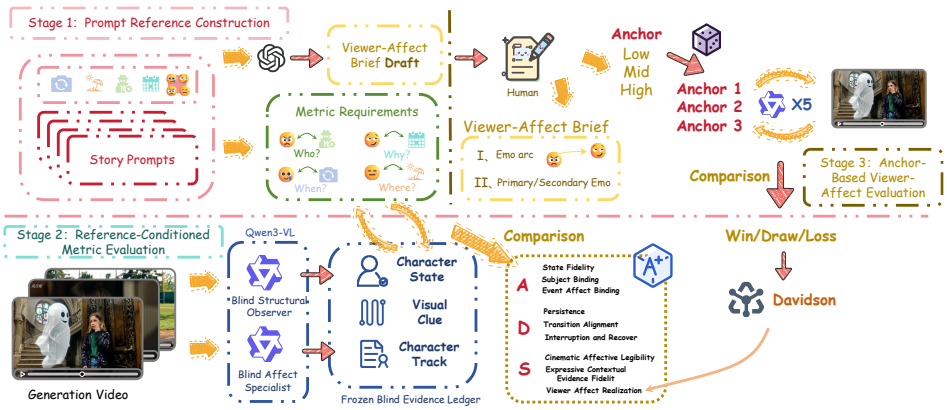


Figure 2: Overview of AffectVBench Eval, covering Accuracy, Development, and Support with evidence extraction, verification, and uncertainty aware scoring for fine grained affective evaluation.

Affective faithfulness in long-form videos cannot be adequately characterized by a single global score, which may obscure qualitatively distinct failure modes. *The intended affect may be expressed by the wrong character or misaligned with the relevant narrative event, fail to persist and evolve*

across shots, or lack sufficiently clear visual evidence to be reliably perceived. To diagnose these failures, we first introduce a common evaluation protocol shared across multiple metrics (Sec. 4.1), and then assess affective faithfulness along three complementary dimensions. **Accuracy** evaluates whether the intended affect is expressed by the correct character and appropriately aligned with the corresponding narrative event (Sec. 4.2). **Development** examines whether the affect is sustained over time and how it evolves throughout the narrative (Sec. 4.3). **Support** assesses whether the intended affect is grounded in sufficient and perceptible visual evidence (Sec. 4.4). Each dimension comprises three metrics, resulting in a nine-metric evaluation suite, as illustrated in Figure 2.

4.1 Common Evaluation Protocol

Blind Evidence Extraction. Evaluation begins with blind evidence extraction, fully separated from the Reference. Qwen3-VL-8B (Yang et al., 2025) Instruct serves as an Affective Forensic Auditor and observes only the generated video, extracting characters, events, actions, temporal relations, and affect related cues into a temporally grounded Evidence Ledger. Grounding-DINO (Liu et al., 2024) and SAM2 (Ravi et al., 2025) are used for character localization and tracking, SFace (Zhong et al., 2021) assists identity verification across shots, feat extracts Action Units, gaze, and head pose, and Keypoint R-CNN (He et al., 2017) and Swin3D-T (Liu et al., 2022) provide body pose and action evidence, respectively. Shot boundaries and motion cues further support temporal analysis. These specialized models provide verifiable evidence rather than directly determining the final affect score.

Reference Conditioned Scoring. The overall evaluation follows a reference conditioned scoring process: $V_i \rightarrow \mathcal{E}_i \rightarrow (\mathcal{E}_i, R_i^m) \rightarrow \text{Verification} \rightarrow S_i^m$, where $\mathcal{E}_i$ denotes the blind Evidence Ledger extracted from the generated video, R_i^m is the metric specific Evaluation Reference, and S_i^m is the resulting score for metric m. Verification is invoked only when the available evidence remains ambiguous. Given $\mathcal{E}_i$ and R_i^m, Qwen3-VL-32B (Yang et al., 2025) Instruct serves as the Reference Conditioned Judge and receives only the metric specific Reference and the relevant Evidence Ledger entries. We use the 32B variant because this stage requires fine grained reasoning over structured reference requirements and temporally grounded visual evidence. When character identity, event localization, or evidence visibility remains uncertain, InternVL3.5-8B (Wang et al., 2025) Instruct independently reinspects the corresponding video interval and returns supplementary evidence to the original judge, which then performs the final assessment.

For each applicable requirement r in State Fidelity, Subject Binding, Event Affect Binding, Persistence, Transition Alignment, Interruption and Recovery, Expressive Contextual Evidence Fidelity, and Cinematic Affective Legibility, the judge predicts a probability distribution over five fulfillment levels: $\mathbf{p}_r = [p_r^0, \ldots, p_r^4] \qquad \sum_{k=0}^{4} p_r^k = 1$. The expected fulfillment is then mapped to a score in $[0, 100]$ where $s_r = 25 \sum_{k=0}^{4} k\, p_r^k$. The full probability distribution is retained to represent evaluator uncertainty rather than collapsing each judgment into a single discrete label. Requirements not specified by the Story Prompt are excluded from scoring. Finally, dependencies among metrics are further handled by predefined error ownership and blocking rules, preventing the same underlying failure from being repeatedly penalized across multiple metrics.

4.2 Emotional Accuracy

Existing affective evaluations typically focus on coarse emotion recognition or global text video correspondence, but do not determine whether a fine grained affective state is expressed by the correct character and appropriately grounded in the relevant narrative event. We therefore decompose Emotional Accuracy into State Fidelity, Subject Binding, and Event Affect Binding, corresponding to *what* affect is expressed, *who* expresses it, and *why* it occurs.

State Fidelity (SF). Rather than treating affect evaluation as closed-set emotion classification, SF formulates affect evaluation as an open vocabulary grounding problem. The target may be fine-grained, intensity sensitive, or compositional, such as *restrained anxiety* or *uneasy relief*, whose meaning cannot be faithfully reduced to a basic emotion category. Let R_i^{SF} denote the SF requirements for sample i. Each requirement is represented as $r = (c_r, \mathcal{A}_r, \lambda_r, \rho_r, \tau_r) \in R_i^{\text{SF}}$, where c_r is the target character, $\mathcal{A}_r$ contains one or more open vocabulary affective states, λ_r specifies affect

intensity when applicable, ρ_r describes relations among multiple states, and τ_r denotes the relevant narrative interval. Evidence associated with the trajectory of c_r over τ_r is retrieved from the blind Evidence Ledger and compared with the requirement. Closed-set emotion predictions are used only as auxiliary cues and never define the target affect.

Subject Binding (SB). Correct affect alone is insufficient if it is expressed by the wrong character, a failure that conventional emotion recognition largely overlooks. SB therefore evaluates whether the observed affect is correctly bound to its intended subject. We represent a requirement as $r = (c_r, \mathcal{A}_r, \tau_r) \in R_i^{\mathrm{SB}}$, and retrieve identity consistent evidence for c_r across the relevant shots. The evaluator then determines whether the observed affective evidence remains associated with the intended character throughout τ_r. If the correct affect is instead expressed by another character, the failure is attributed primarily to SB rather than redundantly penalized under SF.

Event Affect Binding (EAB). Existing temporal alignment metrics may determine whether an event occurs, but not whether the intended character exhibits the required affect *in relation to* that event. EAB explicitly evaluates this event–affect association. We represent each requirement as $r = (c_r, e_r, \mathcal{A}_r, \rho_r) \in R_i^{\mathrm{EAB}}$, where c_r, e_r, and $\mathcal{A}_r$ denote the character, narrative event, and target affect, while $\rho_r \in \{\mathrm{before}, \mathrm{during}, \mathrm{after}\}$ specifies their expected narrative relation. EAB evaluates whether the correct character exhibits the required affect in relation to the intended event, using event relative ordering rather than absolute temporal overlap with the source video.

4.3 Emotional Development

Temporal consistency does not imply affective continuity. A character may remain visually consistent across shots while its affect disappears, changes in the wrong direction, or fails to recover after an interruption. AffectVBench therefore models affect as a narrative trajectory rather than a collection of independent frame level states. Emotional Development evaluates this trajectory through Persistence, Transition Alignment, and Interruption and Recovery.

Persistence (PER). Frame-level emotion recognition provides little insight into whether an established affect remains active across a narrative interval. PER evaluates whether the required affect persists for as long as implied by the Story Prompt and whether it disappears only when narratively justified. For a requirement r, let its narrative interval be partitioned into K windows and let q_k denote the evidence support for the target affect in window w_k. We define $C_{\mathrm{pers}}(r) = \frac{1}{K}\sum_{k=1}^{K} q_k$. The evaluator additionally checks for unsupported resets across shots or events. PER therefore measures long range affective continuity rather than local expression similarity.

Transition Alignment (TA). Correctly detecting two affective states does not guarantee that the transition between them is narratively appropriate. TA evaluates whether the affect changes toward the intended state, in the intended order, and in relation to the intended event. A transition requirement is represented as $\mathcal{T}_r = (a_s, e_r, a_t, \eta_r)$, where a_s and a_t are the source and target affects, e_r is the associated event, and η_r specifies the transition pattern when applicable. The evaluator verifies the source state, target state, transition order, and event-relative timing, and additionally distinguishes gradual, sudden, delayed, or oscillatory transitions when specified.

Interruption and Recovery (IR). Many long-form affective trajectories cannot be captured by a single state transition. An established affect may be temporarily suppressed, replaced, and later recovered. IR therefore represents multistage development as $\mathcal{Z}_r = [z_1, \ldots, z_K]$, with a typical structure $S_{\mathrm{source}} \rightarrow S_{\mathrm{interrupt}} \rightarrow S_{\mathrm{recover}}$. The evaluator verifies both the realization and narrative ordering of these stages. IR consequently captures structured interruption and recovery rather than reducing the trajectory to independent emotion predictions.

4.4 Emotional Support

Correct affective semantics alone do not guarantee successful affective communication. The intended state must be supported by observable behavior, remain perceptually legible under cinematic presentation, and ultimately evoke the intended response in the viewer. We therefore separate Emo-

tional Support into Expressive Contextual Evidence Fidelity, Cinematic Affective Legibility, and Viewer Affect Realization.

Expressive Contextual Evidence Fidelity (ECEF). A model may capture the correct affective semantics without generating the visual evidence through which that affect is conveyed. ECEF evaluates whether the generated video provides sufficient observable support for each affective requirement, including facial expression, gaze, head movement, posture, action, interaction, object relation, and environmental context. Importantly, ECEF evaluates *functional* affective evidence rather than literal source reconstruction. For example, hesitation may be conveyed through a pause, a backward step, or repeated glances, provided that the generated cue fulfills the same affective function.

Cinematic Affective Legibility (CAL). The presence of affective evidence does not imply that it is perceptually accessible to the viewer. CAL evaluates whether such evidence remains readable under the generated cinematography, considering factors such as subject scale, occlusion, blur, illumination, shot scale, focus, and camera motion. A correct expression that is too small, heavily occluded, or obscured by rapid motion may therefore satisfy ECEF but fail CAL. This distinguishes affective legibility from general aesthetic quality.

Viewer Affect Realization (VAR). Character affect and viewer affect are fundamentally different evaluation targets: $\text{Expressed Affect} \neq \text{Viewer Affect}$. Existing automatic evaluation commonly estimates viewer response through a single absolute score, which can be unstable for subjective, open vocabulary affective targets. VAR instead introduces a calibrated relative evaluation protocol. For each target viewer affect, we construct a human verified Anchor Bank spanning low, medium, and high realization levels. A generated video V_i is blindly compared with the corresponding anchors: $V_i \succ A_j, \quad V_i \prec A_j, \quad V_i \sim A_j$. To avoid forcing ambiguous comparisons into binary decisions, ties are explicitly modeled using the Davidson formulation:

$$P(i \succ j) = \frac{e^{u_i}}{e^{u_i} + e^{u_j} + \nu e^{(u_i+u_j)/2}}, \qquad P(i \sim j) = \frac{\nu e^{(u_i+u_j)/2}}{e^{u_i} + e^{u_j} + \nu e^{(u_i+u_j)/2}}, \tag{1}$$

where u_i represents the latent realization strength of the target viewer affect and ν captures tie propensity. The calibrated anchors establish a shared reference scale, after which the inferred utility is mapped to $[0, 100]$.

4.5 Uncertainty Aware Score Aggregation

Affective evidence is inherently ambiguous: the same expression may support several nearby interpretations, while occlusion, weak visibility, and gradual transitions can make a deterministic label unreliable. AffectVBench therefore preserves evaluator uncertainty at the requirement level rather than forcing each judgment into a single discrete category.

For each applicable requirement r in the first eight metrics, the Reference Conditioned Judge predicts a distribution over five fulfillment levels: $\mathbf{p}_r = [p_r^0, \ldots, p_r^4], \quad \sum_{k=0}^{4} p_r^k = 1.$

Its expected fulfillment is mapped to $[0, 100]$ as $s_r = 25 \sum_{k=0}^{4} k\, p_r^k$. Thus, ambiguous evidence contributes proportionally rather than being prematurely collapsed into a hard decision.

For metric m, only applicable requirements are aggregated:

$$S_i^m = \frac{1}{|\mathcal{R}_{i,m}^+|} \sum_{r \in \mathcal{R}_{i,m}^+} s_r. \tag{2}$$

Requirements not authorized by the Story Prompt are excluded, while predefined error ownership and blocking rules prevent one underlying failure from being repeatedly counted across correlated metrics. Metric scores are subsequently macro averaged within Accuracy, Development, and Support, and the overall Affective Faithfulness score is obtained by equally averaging the three dimensions. We report both the dimension scores and the full nine metric profile: $\mathbf{S}_i = [\text{SF}, \text{SB}, \text{EAB}, \text{PER}, \text{TA}, \text{IR}, \text{ECEF}, \text{CAL}, \text{VAR}]$, rather than reducing affective performance to a single scalar.

Table 1: **Overall results on AffectVBench.** General video quality and Affective Faithfulness are reported for 13 T2V systems. ER denotes HERD's *Emotional Response* sub-dimension (Zheng et al., 2025). Affective metrics are normalized to $[0, 100]$. **Bold** and underlined values denote the best and second-best results.

Method	General				Accuracy			Development			Support			Overall
	PQ	OA	DQ	ER	SF	SB	EAB	PER	TA	IR	ECEF	CAL	VAR	
Rolling-Forcing (Liu et al., 2025)	79.0	50.2	59.2	28.3	60.8	66.4	61.5	33.5	32.7	33.8	67.1	68.6	43.5	52.0
Reward-Forcing (Lu et al., 2025)	76.3	50.7	60.5	29.7	61.5	68.2	63.8	28.5	27.9	30.1	67.9	70.4	39.7	50.9
Deep Forcing (Yi et al., 2025)	77.2	51.5	58.6	29.3	61.6	68.9	62.7	32.0	30.8	30.4	68.1	70.3	45.4	52.2
SANA-Video (Chen et al., 2025)	**85.3**	29.0	58.0	25.5	63.3	67.7	62.7	30.2	21.5	22.4	63.4	65.5	27.6	47.1
Infinity-RoPE (Yesiltepe et al., 2026)	80.8	38.3	56.9	29.7	**68.7**	72.1	**69.6**	49.1	50.8	53.4	67.3	68.8	45.6	60.6
MemRoPE (LongLive) (Kim et al., 2026)	80.2	49.9	59.8	28.9	61.4	67.1	63.4	29.4	27.3	27.9	66.9	69.3	44.6	50.8
MemRoPE (Self-Forcing) (Kim et al., 2026)	78.9	50.0	59.2	29.4	63.3	68.3	65.4	31.5	31.0	31.8	65.7	67.8	44.8	52.2
Causal-Forcing (Zhu et al., 2026)	76.5	50.1	53.9	28.7	61.1	65.3	61.8	33.0	31.1	31.8	64.7	66.7	48.6	51.6
Context-Forcing (Chen et al., 2026)	80.9	47.5	59.6	26.8	67.3	**72.2**	68.8	44.2	40.2	40.7	**70.4**	**72.5**	46.6	58.1
Helios (Yuan et al., 2026)	62.9	40.1	**61.2**	26.5	61.2	67.2	61.6	32.1	32.5	34.4	64.5	66.3	37.7	50.8
Stream-R1 (Wu et al., 2026)	77.7	48.1	60.2	29.3	62.3	69.0	65.4	26.1	25.8	26.7	67.1	69.7	41.5	50.4
Seedance 1.0 Pro (Gao et al., 2025b)	77.6	63.9	58.4	**38.2**	63.7	68.9	66.1	55.8	55.3	59.1	63.9	67.4	**57.0**	61.9
Vidu Q2 Pro (Bao et al., 2024)	69.1	**68.9**	50.3	37.9	66.3	69.8	68.1	**59.0**	**57.1**	**61.7**	67.7	69.9	51.3	**63.4**

5 EXPERIMENTS

AffectVBench evaluates 13 representative T2V systems spanning multi-prompt generation, direct long video generation, and proprietary video generation systems. All methods are evaluated using the same Story Prompts and target durations; direct input models receive the complete prompt, while methods requiring multiple prompts use a deterministic decomposition based on narrative events without introducing additional semantic information. We report the general video metrics adopted from LoCoT2V (Zheng et al., 2025), including Perceptual Quality (PQ), Overall Alignment (OA), Dynamic Quality (DQ), and Emotional Response (ER), together with the nine affective metrics defined in Sec. 4. Source videos are evaluated under the same protocol as a *Real Video Reference*. Detailed model descriptions, generation configurations, implementation settings, and evaluation protocols are provided in the Appendix.

5.1 BENCHMARK RESULTS

Table 1 shows that general video quality is not a reliable proxy for affective faithfulness. SANA-Video achieves the highest PQ score (85.3), but performs poorly on TA (21.5) and IR (22.4), yielding the lowest Overall score (47.1). Here, TA measures whether affect changes in the intended direction and order at the appropriate time relative to the associated narrative event, while IR evaluates whether an affective state is correctly interrupted and subsequently recovered. Conversely, Vidu Q2 Pro (Bao et al., 2024) obtains a considerably lower PQ score (69.1) but achieves the highest Overall score (63.4). Similar discrepancies can be observed for models with competitive DQ or OA scores but only moderate affective performance. These results indicate that perceptual quality and coarse prompt alignment mainly capture whether a video looks plausible and depicts the requested content, but do not determine whether characters exhibit the correct event-grounded affective trajectory.

A clear capability gap also emerges across the affective dimensions. Accuracy averages 65.5 across models, whereas Development averages only 36.7, suggesting that current systems can often generate plausible local affective states and associate them with the correct subject, but struggle to sustain, transform, and recover these states across events and shots. In particular, many forcing-based baselines obtain Development scores of only approximately 25–33, indicating that affective cues frequently appear as isolated expressions rather than as coherent narrative progressions. Support exhibits a related asymmetry: ECEF and CAL reach average scores of 66.5 and 68.7, respectively, while VAR remains substantially lower at 44.1, and no evaluated model exceeds 57.0 on this metric. Thus, generating visible and cinematically legible affective evidence does not guarantee that the evidence collectively evokes the intended viewer response. The rankings further demonstrate the diagnostic value of this decomposition: Infinity-RoPE performs best on Accuracy, Context-Forcing leads ECEF and CAL, whereas Vidu Q2 Pro and Seedance 1.0 Pro obtain their overall advantage primarily from stronger Development. Overall, no model dominates all nine metrics, highlighting the importance of evaluating affective faithfulness beyond general video quality or a single coarse emotion score.

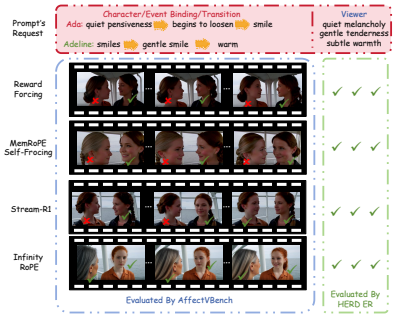


Figure 3: **AffectVBench Eval Overview.** It evaluates Accuracy, Development, and Support with fine-grained affective metrics.

Dimension	Metric	PLCC	SRCC	KRCC	Human Agr.
Accuracy	SF	0.78	0.76	0.59	0.82
	SB	0.84	0.82	0.66	0.86
	EAB	0.72	0.70	0.53	0.78
Development	PER	0.63	0.60	0.45	0.71
	TA	0.69	0.66	0.50	0.74
	IR	0.57	0.54	0.40	0.66
Support	ECEF	0.72	0.69	0.53	0.77
	CAL	0.80	0.78	0.62	0.83
	VAR	0.61	0.58	0.43	0.70

Table 2: **Human alignment of AffectVBench.** Correlations between automatic scores and human judgments across nine affective metrics. Human agreement is reported as a reference.

5.2 Beyond Viewer Response: Affective Narrative Faithfulness

As shown in Figure 3, LoCoT2V introduces Emotional Response (ER) in HERD to assess whether a generated video elicits viewer emotions consistent with the textual expectation. However, ER primarily focuses on the overall emotional impact of the video, i.e., what the viewer ultimately feels, without explicitly modeling who experiences which emotion, what event triggers it, and how it persists or evolves over time. As illustrated in the figure, even when different models successfully convey overall viewer responses such as quiet melancholy, gentle tenderness, or subtle warmth, they may still exhibit clear failures in character binding, event affect association, and affective transitions.

In contrast, AffectVBench extends affective evaluation from a single viewer response into a more complete notion of affective narrative faithfulness. Accuracy evaluates whether the intended affective state is correctly realized and bound to the appropriate character and narrative event. Development assesses whether affect persists across shots, evolves consistently with narrative events, and follows more complex patterns such as interruption and recovery. Support further examines whether the intended affect is supported by sufficient and perceptually legible visual evidence, and whether it ultimately induces the expected viewer response. AffectVBench therefore evaluates not only whether the viewer feels the intended emotion, but also whether that emotional effect is narratively realized through the correct character, event, and temporal trajectory, rather than arising merely from a broadly matching overall atmosphere.

Human Alignment. As shown in Table 2, AffectVBench achieves consistent agreement with human judgments across all nine metrics, with average PLCC, SRCC, and KRCC values of 0.71, 0.68, and 0.52, respectively. Alignment is strongest for visually grounded metrics: Subject Binding (SB) achieves the highest correlation (0.84 PLCC and 0.82 SRCC), followed by Cinematic Affective Legibility (CAL) with 0.80 PLCC and 0.78 SRCC. In contrast, Interruption and Recovery (IR) and Viewer Affect Realization (VAR) show relatively lower correlations, reflecting the greater ambiguity of multi-stage affective trajectories and subjective viewer responses. Notably, this trend is consistent with human agreement, which is also lower for IR and VAR. Overall, these results demonstrate that AffectVBench reliably captures human-perceived affective faithfulness while revealing the remaining difficulty of evaluating long-range affect development and viewer-level affect.

6 Conclusion

This paper has introduced **AffectVBench**, a fine-grained benchmark for evaluating affective faithfulness in long-form T2V generation. Grounded in real videos, AffectVBench combines multi-agent extraction and human verification to construct Story Prompts and structured references capturing open-vocabulary affective states and their temporal evolution. Its nine metrics across Accuracy, Development, and Support, together with blinded anchor-based calibration and uncertainty-aware probabilistic scoring, enable evidence-grounded affective evaluation. Extensive experiments on 13 representative T2V models show that current models still struggling with character–affect binding, cross-shot persistence, and event-driven transitions. AffectVBench thus provides a rigorous foundation for advancing more credible and emotionally coherent long-form video generation.

AI USE STATEMENT

Generative AI tools were used in this work both as components of the proposed benchmark pipeline and as research assistants. In benchmark construction, multimodal foundation models, including Seed2.1 Pro, Qwen3.8 Max, MiniMax M3, GPT-5.6 Sol, and Seed2.1 Turbo, were used to assist with audiovisual narrative understanding, muted visual verification, cross-model consistency checking, structured metadata construction, Story Prompt generation, and reference construction. Generative and vision-language models were also used as components of the evaluation pipeline for evidence reasoning and reference-conditioned assessment, as described in the main paper. All AI-assisted benchmark artifacts were subject to predefined validation procedures and human review, and the final benchmark specifications and annotations were determined by the authors rather than directly accepting model outputs.

Generative AI tools were additionally used to assist with research methodology refinement, code development and debugging, data processing scripts, experimental analysis, scientific figure preparation, translation, and drafting or editing portions of the manuscript. The authors manually reviewed AI-assisted text, code, analyses, and artifacts, and experimental outputs were checked against the underlying data and evaluation records. AI tools were not used to fabricate experimental measurements, replace required human verification, or autonomously determine the final scientific conclusions of the paper. We take full responsibility for the final content of this work, including all text, claims, code, data artifacts, and experimental results produced with the aid of generative AI.

REFERENCES

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Lancheng Gao, Ziheng Jia, Yunhao Zeng, Wei Sun, Yiming Zhang, Wei Zhou, Guangtao Zhai, and Xiongkuo Min. Eemo-bench: a benchmark for multi-modal large language models on image evoked emotion assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7064–7073, 2025a.

Lancheng Gao, Ziheng Jia, Shengyan Li, Zixuan Xing, Jiarui Wang, Huiyu Duan, and Xiongkuo Min. E3 mo-bench: A scalable benchmark for multimodal evoked and expressed emotion understanding via bayesian pairwise alignment. *arXiv preprint arXiv:2608.10796*, 2026.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025b.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

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

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2881–2889, 2020.

Youngrae Kim, Qixin Hu, C-C Jay Kuo, and Peter Beerel. Memrope: Training-free infinite video generation via evolving memory tokens. In *European Conference on Computer Vision*, pp. 244–261. Springer, 2026.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, et al. Minimax sparse attention. *arXiv preprint arXiv:2606.13392*, 2026.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations*, volume 2026, pp. 23406–23432, 2026.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3192–3201. IEEE, 2022.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. Veatic: Video-based emotion and affect tracking in context dataset. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4455–4465. IEEE, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, et al. Filmbench: A film-grade benchmark for cinematic video generation. *arXiv preprint arXiv:2607.24241*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Bin Wu, Mengqi Huang, Shaojin Wu, Weinan Jia, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Stream-r1: Reliability-perplexity aware reward distillation for streaming video generation. *arXiv preprint arXiv:2605.03849*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2236–2246, 2018.

AmirAli Bagher Zadeh, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1801–1812, 2020.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, YiFan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 48767–48807, 2026.

Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. Locot2v-bench: Benchmarking long-form and complex text-to-video generation. In *Forty-third International Conference on Machine Learning*, 2025.

Yaoyao Zhong, Weihong Deng, Jiani Hu, Dongyue Zhao, Xian Li, and Dongchao Wen. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 30:2587–2598, 2021.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Table 3: **Comparison between Raw Prompt and AffectVBench on Causal-Forcing.** The table reports eight affective metrics across Accuracy, Development, and Support. Δ denotes the signed difference, computed as AffectVBench minus Raw Prompt.

Dimension	Metric	Raw Prompt	AffectVBench	Δ
Accuracy	SF	38.93	61.08	+22.15
	SB	39.34	65.28	+25.94
	EAB	37.49	61.78	+24.29
Development	PER	39.11	33.03	−6.08
	TA	28.08	31.11	+3.03
	IR	30.54	31.77	+1.23
Support	ECEF	56.67	64.67	+8.00
	CAL	54.50	66.71	+12.21
Mean		40.58	51.93	**+11.35**

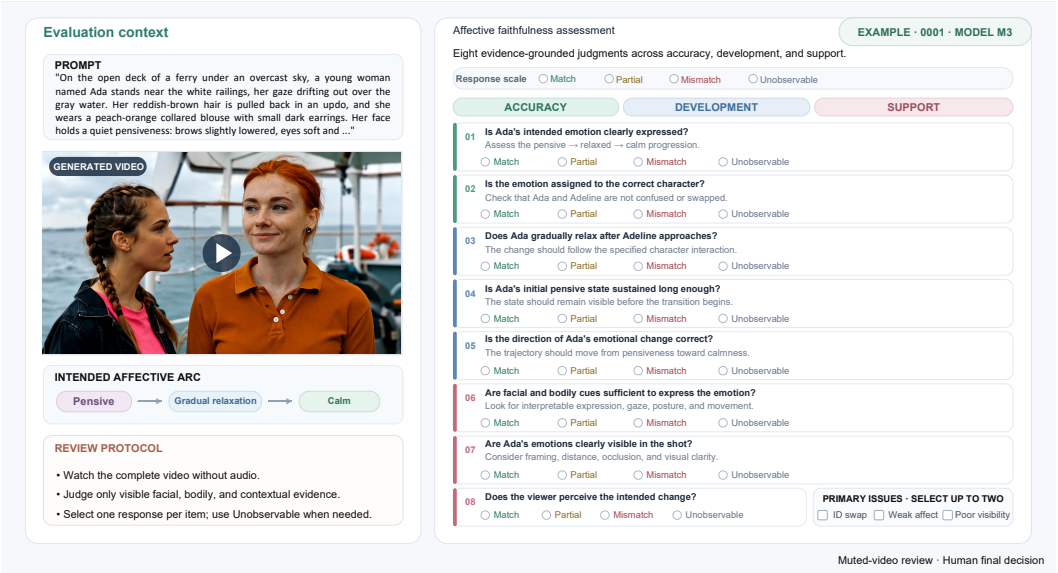


Figure 4: **Human evaluation questionnaire.** The questionnaire used to collect human judgments for AffectVBench.

A Prompt Construction Ablation

To isolate the contribution of AffectVBench's Story Prompt construction, we compare it with the original raw prompt on Causal-Forcing while keeping the generation and evaluation protocols fixed. The Raw Prompt condition uses the original input prompt, whereas the AffectVBench condition uses the corresponding human-verified Story Prompt. Since Viewer Affect Realization (VAR) relies on a separate anchor-based calibration protocol, this comparison covers the other eight affective metrics.

As shown in Table 3, AffectVBench substantially improves affective accuracy and visual support. The largest gains occur in Subject Binding (SB, $+25.94$), Event–Affect Binding (EAB, $+24.29$), and State Fidelity (SF, $+22.15$), indicating that the Story Prompt better specifies affective states, character attribution, and event grounding. Support also improves, particularly in Cinematic–Affective Legibility (CAL, $+12.21$). In contrast, Persistence (PER) decreases by 6.08, suggesting that improved prompt specificity alone does not guarantee long-range affective continuity. Overall, the mean score increases from 40.58 to 51.93, corresponding to an 11.35-point improvement across the eight directly scored metrics.

B Questionnaire Design and Sampling

Figure 4 shows the questionnaire used for human evaluation.

C Experimental Settings

C.1 Baselines and Generation Protocol

We evaluate 13 representative T2V systems, including 11 open-source systems and two proprietary systems. The open-source systems are Rolling-Forcing (Liu et al., 2025), Reward-Forcing (Lu et al., 2025), Deep Forcing (Yi et al., 2025), SANA-Video (Chen et al., 2025), Infinity-RoPE (Yesiltepe et al., 2026), MemRoPE with LongLive (Kim et al., 2026), MemRoPE with Self-Forcing (Kim et al., 2026), Causal-Forcing (Zhu et al., 2026), Context-Forcing (Chen et al., 2026), Helios (Yuan et al., 2026), and Stream-R1 (Wu et al., 2026). The two MemRoPE configurations are treated as separate systems because they use different base models while sharing the same MemRoPE memory-token mechanism. The proprietary systems are Seedance 1.0 Pro (Gao et al., 2025b) and Vidu Q2 Pro (Bao et al., 2024).

Except for Infinity-RoPE, all open-source systems receive the complete AffectVBench Story Prompt following their native inference procedures. We do not perform model-specific semantic rewriting or prompt simplification for these systems. Infinity-RoPE requires multiple prompt inputs, therefore, each Story Prompt is decomposed into an ordered sequence of verbatim segments using the generate–then–verify procedure described in Sec. C.2.1. The same decomposition is used for all Infinity-RoPE samples. The segmentation first generates a conceptual scene-boundary reference, then extracts exact consecutive substrings from the frozen Story Prompt, followed by programmatic reconstruction and semantic quality checks. Thus, concatenating all final segments reconstructs the original Story Prompt character by character.

All generated videos target a resolution of 480p and a duration of approximately 30 seconds. Audio generation is disabled for all systems, and only the visual stream is used for evaluation. Models whose native inference pipeline directly supports the target duration generate the complete video according to their original long-video procedure. For systems with a native generation duration shorter than 30 seconds, we use a segment-wise autoregressive generation protocol. Specifically, the Story Prompt is divided into three consecutive temporal segments, each targeting approximately 10 seconds. The first segment is generated from text. Each subsequent segment is conditioned on its corresponding segment prompt and the last frame of the preceding segment. The generated segments are then temporally concatenated to form the final video. This protocol is used for Seedance 1.0 Pro and Vidu Q2 Pro because their native generation interfaces do not directly produce a single 30-second clip.

C.2 Prompts Used in Our Pipeline

We provide the prompts used in our AffectVBench multi-prompt construction pipeline. The prompts are organized into three stages: prompt segmentation, exact segmentation and textual verification, and visual post-checking. All prompts are shown verbatim.

C.2.1 Prompt Segmentation

Text Segmentation Prompt

```
You are an expert in affect-aware cinematic scene segmentation and multi-prompt video
generation for AffectVBench.

Your task is to divide a frozen long-form visual Story Prompt into an ordered sequence
of video-generation-friendly conceptual segments. The segments must preserve the
prompt's complete person-event-time-perception structure and affective arc. The Story
Prompt is the sole source of truth; do not add, remove, correct, or infer story content.

Boundary rules:
```

```
1. Create a boundary at a clear physical scene, location, shot sequence, or major time
change.
2. Within the same physical space, allow a boundary when necessary to preserve a
distinct scorable affective phase: affect onset, event-triggered response, persistence
interval, state or intensity change, transition milestone, interruption, suppression,
inserted state, recovery, re-escalation, subject-binding change, or cinematic-legibility
change.
3. Do not split merely at every sentence, camera detail, or compatible cue within one
stable phase.
4. Each segment should contain one primary scene/action-affect phase and at most one or
two tightly related sub-actions. Keep an event, the affected character, and its
immediate visible reaction together when separating them would destroy event-affect
binding.
5. Preserve continuous affect persistence without implying false disappearance, reset,
or repeated onset.
6. Preserve transition direction, order, gradual-versus-abrupt type, relative event
timing, and intensity trajectory.
7. Preserve interruption-and-recovery stages without merging away any explicitly stated
stage.
8. Preserve correct character identity and affect binding. Do not create ambiguous
pronoun-only segments.
9. Preserve all stated appearance, clothing, objects, interactions, environment, visible
expression, gaze, posture, action, and cinematic cues exactly in meaning.
10. Preserve stated event-affect causal or temporal relations; never infer a cause from
mere succession.
11. Do not invent visible cues for abstract emotions and do not infer viewer affect.
12. Keep fewer than 20 segments and prefer the fewest segments that preserve physical
and affective stages.
13. Do not add timestamps, evidence notes, confidence labels, IDs, or model-specific
repeated anchors.
14. Preserve the full semantic content and chronological order of the original prompt.

Output only:
<list>["prompt1", "prompt2", ..., "..."]</list>

Frozen canonical Story Prompt:
{prompt_text}
```

Video Evidence Prompt

```
You are the visual evidence observer for AffectVBench multi-prompt construction.

You are given a frozen canonical Story Prompt and a chronological contact sheet sampled
from its corresponding source movie clip. The contact sheet is muted visual evidence
only. Each frame is labeled with a normalized timestamp from 0.000 to 1.000.

Your task is to report evidence that may help locate physical, event, and affective
phase boundaries. This is an evidence-extraction step, not a prompt-writing step.

Rules:
1. Treat the frozen Story Prompt as the sole authority for generation semantics. The
video may clarify where an already stated character, event, affect, transition, or
visible cue occurs, but it must not add a new requirement.
2. Do not import dialogue, subtitles, audio, music, sound effects, character names,
backstory, causal claims, or emotions that are absent from the frozen Story Prompt.
3. Report only visible evidence from the provided frames. Distinguish direct observation
from uncertainty caused by sparse sampling, blur, occlusion, distance, or an ambiguous
expression.
4. Track which distinguishable character displays each affective cue. Do not transfer a
cue between people.
5. Preserve the difference between character affect and viewer affect. Do not infer
viewer affect.
6. Identify approximate normalized ranges for physical scenes, major events, stable
affective phases, transition milestones, interruptions, recoveries, and changes in
cinematic legibility when supported by the frames and the prompt.
7. A sampled frame is not an exact boundary. Report approximate ranges and uncertainty
rather than false precision.
8. If the video visibly conflicts with the frozen prompt, record the conflict for human
review. Do not rewrite the prompt to match the source video.

Return JSON only:
{{
  "physical_phases": [
    {{"start": 0.0, "end": 1.0, "visible_evidence": "", "confidence": "high | medium |
    low"}}
  ],
```

```
  "affective_phases": [
    {{
      "start": 0.0,
      "end": 1.0,
      "character_from_prompt": "",
      "state_or_transition_from_prompt": "",
      "visible_evidence": "",
      "confidence": "high | medium | low"
    }}
  ],
  "event_affect_links": [
    {{"event_from_prompt": "", "reaction_from_prompt": "", "approximate_range": [0.0,
    1.0], "confidence": "high | medium | low"}}
  ],
  "cinematic_legibility_notes": [],
  "prompt_video_conflicts": [],
  "uncertainties": []
}}

Frozen canonical Story Prompt:
{prompt_text}
```

Video-Grounded Segmentation Prompt

```
You are an expert in affect-aware cinematic scene segmentation and multi-prompt video
generation for AffectVBench.

Your task is to divide a frozen long-form visual Story Prompt into an ordered sequence
of video-generation-friendly conceptual segments. Use the supplied muted-video evidence
only to locate boundaries among content already authorized by the Story Prompt. The
segments must preserve the prompt's complete person-event-time-perception structure and
affective arc.

The Story Prompt is semantically authoritative. Video evidence is boundary evidence, not
a source of new generation content. Never add, remove, correct, or replace prompt
content based on the source video.

Boundary rules:
1. Create a boundary at a clear physical scene, location, shot sequence, or major time
change.
2. Within the same physical space, allow a boundary when necessary to preserve a
distinct scorable affective phase: affect onset, event-triggered response, persistence
interval, state or intensity change, transition milestone, interruption, suppression,
inserted state, recovery, re-escalation, subject-binding change, or cinematic-legibility
change.
3. Do not split merely at every sentence, camera detail, or compatible cue within one
stable phase.
4. Each segment should contain one primary scene/action-affect phase and at most one or
two tightly related sub-actions. Keep an event, the affected character, and its
immediate visible reaction together when separating them would destroy event-affect
binding.
5. Preserve continuous affect persistence without implying false disappearance, reset,
or repeated onset.
6. Preserve transition direction, order, gradual-versus-abrupt type, relative event
timing, and intensity trajectory.
7. Preserve interruption-and-recovery stages without merging away any explicitly stated
stage.
8. Preserve correct character identity and affect binding. Do not create ambiguous
pronoun-only segments.
9. Preserve all stated appearance, clothing, objects, interactions, environment, visible
expression, gaze, posture, action, and cinematic cues exactly in meaning.
10. Preserve stated event-affect causal or temporal relations; never infer a cause from
mere succession.
11. Do not invent visible cues for abstract emotions and do not infer viewer affect.
12. Keep fewer than 20 segments and prefer the fewest segments that preserve physical
and affective stages.
13. Do not put timestamps, evidence notes, confidence labels, IDs, or model-specific
repeated anchors into segment text.
14. If video evidence conflicts with the prompt or is too sparse to resolve a boundary,
follow the prompt and choose a conservative textual boundary.

Output only:
<list>["prompt1", "prompt2", ..., "..."]</list>

Frozen canonical Story Prompt:
{prompt_text}
```

```
Muted-video boundary evidence:
{video_evidence}
```

C.2.2 Exact Segmentation and Textual Verification

Verbatim Refinement Prompt

```
You are the verbatim segmentation stage of the AffectVBench multi-prompt construction
pipeline.

Cut the frozen canonical Story Prompt into consecutive exact substrings using the
conceptual segmentation reference only to locate boundaries. The Story Prompt is
authoritative.

Rules:
1. Do not rewrite, paraphrase, summarize, expand, simplify, normalize, translate, or
edit the original text.
2. Every segment must be an exact consecutive substring of the original text.
3. Do not add, remove, duplicate, reorder, capitalize, lowercase, or alter any
character, punctuation, line break, or whitespace.
4. Preserve character identity, affect binding, event-affect relations, affect state and
intensity, visible evidence, transition, persistence, interruption, recovery, and
cinematic instructions exactly as written.
5. Do not add repeated identity, appearance, clothing, environment, affect, or camera
anchors.
6. If a conceptual boundary cannot be represented exactly, use the nearest exact cut
that best preserves its physical or affective phase.
7. Return 1 to 19 non-empty segments in chronological order.
8. Concatenating all segments with no separator must reconstruct the original prompt
character-for-character.
9. Return only a valid JSON array; add no IDs, timestamps, comments, Markdown, or
explanations.

Original frozen Story Prompt:
{prompt_text}

Conceptual segmentation reference:
{raw_scene_splits}
```

Text Check Prompt

```
You are the independent quality evaluator for AffectVBench multi-prompt segmentation.

Judge whether the segmented list is a benchmark-faithful representation of the frozen
Story Prompt. Do not repair or rewrite it.

Check all of the following:
1. Ordered, consecutive, non-empty exact substrings.
2. Direct concatenation reconstructs the original prompt character-for-character.
3. No addition, omission, duplication, reordering, paraphrase, weakening, strengthening,
normalization, or cross-character transfer.
4. Fewer than 20 segments and no model-specific additions.
5. Physical scenes and major time changes are preserved without sentence-level
over-fragmentation.
6. A1 affect states, optional secondary affects, relations, and intensities remain
complete.
7. A2 affect and visible cues remain bound to the correct distinguishable character
without pronoun ambiguity.
8. A3 events remain correctly related to the affected character's reaction without
invented causality.
9. D1 continuous affect is not made to appear reset or repeatedly reintroduced.
10. D2 transition states, direction, order, type, intensity trajectory, and event timing
remain recoverable.
11. D3 interruption, inserted state, recovery, and re-escalation remain distinct and
ordered when present.
12. S1 facial, gaze, posture, body, behavior, interaction, object, and environmental
cues remain attached to the correct affect phase.
13. S2 explicit framing, focus, lighting, occlusion, and camera cues remain attached to
the evidence they make legible.
14. S3 viewer affect is preserved only if explicitly present and is not confused with
character affect.
```

```
Return JSON only:
{{"Result":"Yes | No","Violated_Rules":[],"Explanation":""}}

Original frozen Story Prompt:
{prompt_text}

Segmented prompt list:
{scene_prompts}
```

C.2.3 VISUAL POST-CHECK

Video Post-check Prompt

```
You are the post-segmentation visual boundary verifier for AffectVBench.

Given the frozen Story Prompt, its exact multi-prompt splits, a chronological
muted-video contact sheet, and previously extracted video evidence, determine whether
the boundaries are visually and affectively plausible. This step may recommend boundary
adjustments only between exact substrings of the frozen prompt. It must not add, remove,
or rewrite semantic content.

Rules:
1. Verify that segment order follows the visible chronology when the video provides
sufficient evidence.
2. Verify physical scene changes, major event boundaries, character-affect binding,
event-affect proximity, affect persistence, transition order, interruption/recovery, and
cinematic legibility.
3. Do not require every camera cut to become a prompt boundary.
4. Do not use source-video content absent from the prompt as a generation requirement.
5. If the prompt and video conflict, flag human review; the video cannot silently change
the frozen prompt.
6. If sparse frames cannot resolve a boundary, return uncertain rather than guessing.
7. Recommend time fractions only as approximate monotonic ranges. The first start must
be 0.0 and final end 1.0. They are scheduling metadata, not text.

Return JSON only:
{{
  "Result": "Yes | No | Uncertain",
  "boundary_findings": [],
  "prompt_video_conflicts": [],
  "recommended_time_fractions": [
    {{"segment_index": 1, "start": 0.0, "end": 1.0, "confidence": "high | medium |
    low"}}
  ],
  "requires_human_review": false,
  "Explanation": ""
}}

Frozen canonical Story Prompt:
{prompt_text}

Exact multi-prompt splits:
{scene_prompts}

Pre-split muted-video evidence:
{video_evidence}
```