EMOTRACE-BENCH: DIAGNOSING EMOTIONAL EXPRESSION IN TEXT-TO-VIDEO GENERATION THROUGH CONTROLLED CUES AND INTENSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-video (T2V) models are increasingly used to generate videos with targeted emotional expressions, yet existing evaluation methods mainly rely on holistic video–text alignment and provide limited insight into whether failures arise from target-emotion expression, intensity control, or specific affective cues. We introduce EmoTrace-Bench, a fine-grained diagnostic benchmark that controls emotion intensity and selected affective cues in prompts and evaluates generated videos at multiple levels. EmoTrace-Bench contains 448 controlled prompts covering eight emotions, two intensity levels, and diverse cue conditions. Using 12 T2V models, we generate 5,376 videos and evaluate target-emotion presence, perceived intensity, and ten audiovisual dimensions of emotional expression with an automated evaluator calibrated and validated against human annotations on a subset of the data. Results reveal substantial differences across models and prompt designs. Prompts that explicitly describe a character's emotional reaction are followed more reliably than prompts that describe only events or scenes, whereas adding more affective cues or using more explicit cinematic audiovisual language does not consistently improve emotional generation. These results illustrate how aggregate alignment scores may overlook localized differences in target-emotion expression, intensity control, and cue realization. EmoTrace-Bench provides a traceable and interpretable framework for diagnosing emotional video generation without modifying existing generation pipelines, and we release the prompts, annotation guidelines, evaluation protocols, and automated scores.

1 INTRODUCTION

Generating videos with explicit emotional expression is an important capability for controllable content generation in text-to-video (T2V) models. For controllable emotion generation, the key question is not only whether a video presents the target emotion, but also whether it follows the prompt's requirements regarding the emotion category, emotion intensity, and specific expressive cues. Emotion generation in T2V is therefore not a single-dimensional problem; rather, it simultaneously involves multiple interrelated yet distinguishable capabilities, including emotion category, emotion intensity, and specific emotional cues.

However, existing T2V evaluations have difficulty effectively distinguishing these capabilities. Existing multidimensional video generation benchmarks mainly focus on semantic consistency, visual quality, motion performance, and compositional relationships, and they generally lack explicit modeling of emotional expression (Huang et al., 2024; 2026; Zheng et al., 2025; Liu et al., 2024; Han et al., 2025). The few works oriented toward emotion generation also mostly use aggregate scores to measure the overall match between generated videos and emotion prompts (Qiu et al., 2026). Although these metrics can indicate whether a video aligns with its prompt overall, they provide limited insight into the specific causes of mismatches and may collapse qualitatively different generation errors into similar aggregate scores, thereby obscuring meaningful differences in models' ability to follow fine-grained emotional instructions. As a result, such evaluations cannot fully characterize a model's controllability over emotion category, intensity, and mode of expression in practical content creation.

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

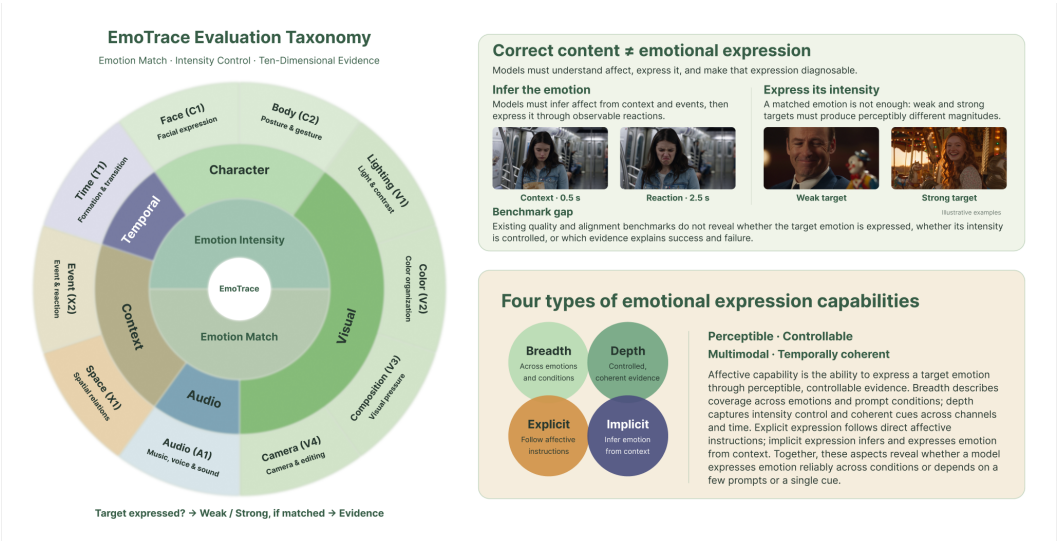


Figure 1: EmoTrace-Bench at a glance. Left: the three-level evaluation taxonomy—whether the target emotion is expressed, whether its perceived intensity is weak or strong, and which of ten expression channels (character, visual form, audio, context, temporal) provide the evidence. Top right: correct content does not guarantee emotional expression; a model must infer the emotion from context and reaction and realize the requested intensity, and existing benchmarks do not diagnose where this fails. Bottom right: the four capability facets examined in Section 4—breadth across emotions and prompt conditions, depth of intensity control and cue support, and explicit versus implicit cue expression.

Adding more output-side metrics alone is insufficient, because confounds often arise during prompt construction. Emotional prompts typically combine multiple cues, making it difficult to attribute output changes to specific factors. Differences across emotion categories can further confound intensity comparisons. Reliable diagnosis therefore requires controlled inputs as well as fine-grained outputs.

We propose EmoTrace-Bench (Figure 1), which treats prompts as controlled experimental variables. It varies two main factors: affective cues and emotion intensity. Cue-controlled prompts change only one cue phrase while keeping the rest fixed, with additional paraphrased variants to test sensitivity to wording. Intensity-controlled prompt pairs differ only in intensity-related terms while preserving the emotion category, scene, and other content. This design enables separate analysis of emotion category, intensity, and specific cues. Figure 2 summarizes the overall pipeline.

At the output, we annotate target-emotion success, perceived intensity, and ten audiovisual cue dimensions. Using 448 controlled prompts, we generate 5,376 videos from 12 T2V models and collect human annotations for 1,120 videos from three of them to validate the automated evaluation. The analysis reveals that emotion-category generation and intensity control are distinct capabilities, model performance can be sensitive to cue wording, and automated evaluation agrees broadly with human judgments while small systematic differences remain on some fine-grained dimensions.

By controlling variables at the input and conducting fine-grained diagnosis at the output, EmoTrace-Bench decomposes emotion-generation performance that was previously mixed together in overall scores into more specific dimensions and prompt factors. At the same time, through per-video, per-attribute comparisons between human and automated evaluation, it further reveals the scope of applicability and limitations of automated evaluation across different emotional dimensions.

The main contributions of this paper are threefold:

(i) We introduce **EmoTrace-Bench**, a fine-grained diagnostic benchmark for emotion generation in text-to-video models. It decomposes emotion generation into three complementary levels: emotion category, emotion intensity, and ten affective cue channels, enabling localized diagnosis of generation failures.

(ii) We develop a controlled prompt-construction framework that independently varies emotion intensity and affective cues under a minimal-difference design. It also includes semantically matched cue variants with different linguistic realizations, allowing us to study prompt sensitivity while reducing confounding across input factors.

(iii) We benchmark 12 text-to-video models on 448 controlled prompts, yielding 5,376 generated videos, and collect human annotations for 1,120 videos to validate the automated evaluation protocol. Our analysis reveals substantial differences across models in emotion-category generation, intensity control, and prompt sensitivity, while showing that aggregate automated judgments are broadly aligned with human annotations. We release the prompts, annotation guidelines, evaluation protocol, and automated scores.

2 Related Work

2.1 Text-to-Video Generation Evaluation

T2V evaluation has evolved from overall quality assessment toward multidimensional diagnosis. VBench, VBench++, VBench-2.0, EvalCrafter, and Video-Bench evaluate visual quality, motion, temporal consistency, subject consistency, spatial relations, text alignment, and higher-level generation fidelity (Huang et al., 2024; 2026; Zheng et al., 2025; Liu et al., 2024; Han et al., 2025). VABench and T2AV-Compass further consider audio quality and audiovisual consistency (Hua et al., 2026; Cao et al., 2025), while T2V-CompBench uses controlled prompts to evaluate compositional generation (Sun et al., 2025).

Despite this progress, these benchmarks mainly target general generation quality, semantic adherence, and compositionality. Emotional expression is rarely modeled as a structured capability with separable dimensions. Existing evaluations therefore provide limited diagnosis of whether errors arise from emotion category, intensity, or specific character, scene, and audiovisual cues.

2.2 Emotional Video Generation and Controllable Expression

Recent work has begun to study emotion-aware video generation. EmoVid provides a multimodal dataset for emotion-centric video understanding and generation (Qiu et al., 2026); FG-ECVG explores fine-grained controllable emotional video generation (Wei et al., 2025); and EmoTalker studies editable emotional expression in talking-face generation (Zhang et al., 2024). EmoWorld further disentangles atmosphere, affective semantic cues, and temporal progression, showing that video emotion contains multiple controllable components (Wang et al., 2026a).

These works mainly focus on emotional data, generation, or model-side control. Their evaluations still commonly rely on emotion-recognition accuracy, condition matching, or aggregate scores. Such metrics may assign similar scores to different failures, such as generating the correct emotion at the wrong intensity or creating an appropriate atmosphere without the requested facial or bodily expression. Thus, an overall emotion score is insufficient for diagnosing fine-grained emotional controllability.

2.3 Fine-Grained Emotional and Audiovisual Cues

Emotion is conveyed through multiple channels. Psychological studies highlight the roles of facial expression, body posture, and movement (Ekman & Friesen, 1978; Cohn et al., 2007; Coulson, 2004; Dael et al., 2012a;b; Aviezer et al., 2012), while film studies show that shot scale, motion, lighting, color, sound, and music also shape emotional perception (Bálint et al., 2018; Detenber et al., 1998; Chion, 1994; Cohen, 2010; Bordwell et al., 2020). Recent benchmarks such as Stable Cinemetrics and FilmBench evaluate structured cinematic attributes (Chatterjee et al., 2025; Wang et al., 2026b), while HVEval and HUG-VIS examine human-centered appearance, action, and emotion (Wu et al., 2025; Ma et al., 2026).

Building on these directions, EmoTrace-Bench decomposes emotion generation into emotion category, emotion intensity, and specific affective cues. It combines controlled prompts with fine-grained output evaluation to diagnose whether the target emotion is generated, whether its intensity follows

Table 1: **Capability coverage of representative video-generation benchmarks.** Filled, dotted, and open circles denote full, partial or domain-restricted, and absent support, respectively.

	Affective target			*Controlled input*			*Diagnostic output*			*Human validation*		
Benchmark	EC	EI	AC	CI	IP	GF	CL	EL	DS	PL	SV	MR
EmoVid (Qiu et al., 2026)	●	○	⊙	○	○	⊙	○	●	●	○	⊙	●
HVEval (Wu et al., 2025)	⊙	○	⊙	○	○	⊙	○	⊙	●	○	⊙	●
HUG-VIS (Ma et al., 2026)	●	○	○	○	○	⊙	○	●	●	○	⊙	●
Stable Cinemetrics (Chatterjee et al., 2025)	○	○	⊙	⊙	○	●	⊙	○	●	○	⊙	⊙
FilmBench (Wang et al., 2026b)	⊙	○	⊙	○	○	⊙	○	⊙	●	○	⊙	●
T2V-CompBench (Sun et al., 2025)	○	○	○	○	○	●	○	○	●	○	●	●
VBench family (Huang et al., 2024; 2026; Zheng et al., 2025)	○	○	○	○	○	●	○	○	●	○	○	●
Video-Bench (Han et al., 2025)	○	○	○	○	○	○	○	○	●	○	●	●
VABench (Hua et al., 2026)	○	○	○	○	○	⊙	○	○	●	○	⊙	●
T2AV-Compass (Cao et al., 2025)	○	○	○	○	○	⊙	○	○	●	○	⊙	●
EvalCrafter (Liu et al., 2024)	○	○	○	○	○	⊙	○	○	●	○	○	●
EmoTrace-Bench (ours)	●	●	●	●	●	●	●	●	●	●	●	⊙

Affective target: EC = emotion category, EI = perceived emotion intensity, AC = fixed affective-cue taxonomy. **Controlled input:** CI = affective-cue isolation, IP = minimally differing intensity pairs, GF = general factor control. **Diagnostic output:** CL = per-video cue-channel labels, EL = per-video emotion labels, DS = dimension- or criterion-level scores. **Human validation:** PL = per-attribute label agreement, SV = sample- or video-level validation, MR = model-ranking validation. Prompt-side control is required for full credit under CI and IP; capture-side factorial designs receive partial credit.

the prompt, and through which character, scene, and audiovisual cues it is expressed. This moves emotion-generation evaluation from holistic matching toward traceable fine-grained diagnosis.

Table 1 summarizes these differences in affective targets, controlled inputs, diagnostic outputs, and human validation.

3 METHOD

3.1 EVALUATION TARGET DESIGN

EmoTrace-Bench evaluates emotional video generation at three levels: target emotion, perceived intensity, and affective cues. We first define the target emotion set and then specify the audiovisual dimensions used for fine-grained diagnosis.

Selection of Eight Emotions. EmoTrace-Bench includes eight emotions: amusement, excitement, awe, contentment, anger, fear, disgust, and sadness. The selection follows the discrete emotion set established by Mikels et al. using the International Affective Picture System (IAPS) (Mikels et al., 2005). It contains four positive and four negative emotions spanning different typical arousal levels (Russell, 1980).

Three-Level Evaluation. For each generated video, the first level determines whether the target emotion is expressed. The second evaluates perceived emotion intensity as strong or weak when the target emotion is present, and as "not matched" otherwise. The third identifies which audiovisual cue dimensions support the target emotion. These levels capture whether the intended emotion is expressed, how strongly it is expressed, and through which cues.

Multidimensional Affective Cues. Emotional expression in video can arise from character behavior, environmental context, and audiovisual form. Drawing on theories of emotional expression, film form, and an external inventory of affective cues (Smith, 2003), we define ten dimensions: facial expression, bodily behavior, lighting, color, composition, visual dynamics, temporal organization, sound, scene, and event. We retain cues that can be directly observed and judged relatively independently in short videos, while excluding factors that require long-form narrative, dialogue semantics, or broader temporal context. Operational definitions and channel-specific prerequisites are provided in Appendix C (Table C.1).

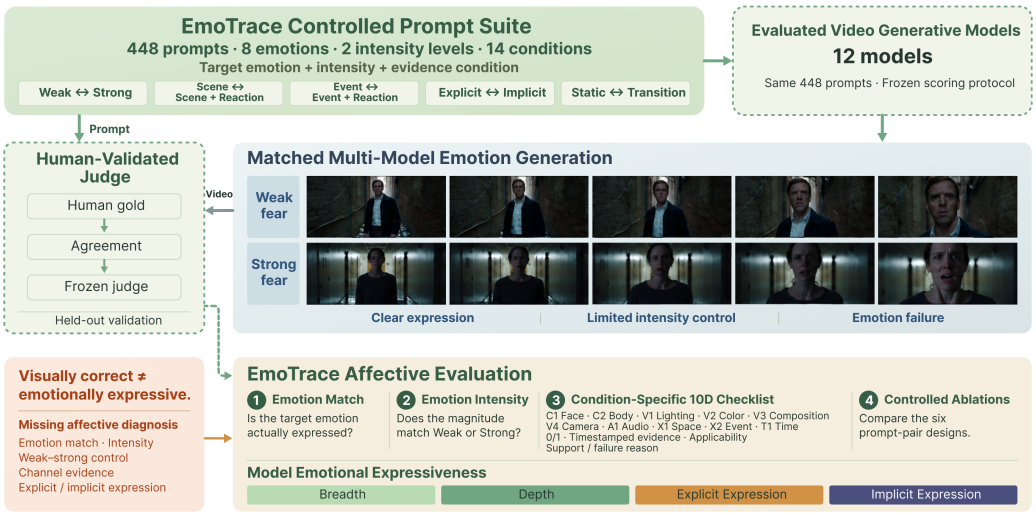


Figure 2: Overview of the EmoTrace-Bench pipeline. Prompts from the controlled suite (top) are sent unchanged to every model; the middle row shows matched weak/strong generations for one fear prompt pair across models. All videos are scored by a single automated evaluator that is frozen after calibration against human labels (left) and applied through the two-stage affective evaluation (bottom).

3.2 Controlled Prompt Set

EmoTrace-Bench uses controlled prompts to vary emotion intensity, affective cues, and prompt formulation while minimizing unrelated changes across conditions.

Scale and Structure. Each prompt is represented as

$$P(e, s, i, t) = B(e, s) + I(e, i) + \sum_{d \in \mathcal{M}_t} C_{d,t}(e, s),$$

where e is the target emotion, s the scene, i the intensity level, and t the prompt template. $B(e, s)$ is the shared base description, $I(e, i)$ specifies emotion intensity, and $C_{d,t}(e, s)$ denotes the realization of affective cue d under template t. The template determines which cue segments are included through $\mathcal{M}_t$ and, for selected conditions, how the same affective content is linguistically expressed.

The benchmark crosses 8 emotions, 2 intensity levels with 2 samples each, and 14 prompt templates, yielding 448 prompts. The templates include emotion-only prompts, individual facial, bodily, scene, and event cues, combinations of character and environmental cues, alternative linguistic realizations, facial changes, empty scenes, and prompts containing all explicitly writable cues. Eight of the ten evaluation dimensions—facial expression, bodily behavior, lighting, color, composition, visual dynamics, scene, and event—can be explicitly specified in prompts. Temporal organization and sound are evaluated only in the generated video. Each prompt records whether a cue dimension is explicitly specified or left unspecified but observable, allowing explicit instruction following to be distinguished from cues that emerge without being requested. Representative prompt conditions are shown in Figure B.1 in Appendix B.

Alternative Linguistic Realizations. To examine sensitivity to prompt formulation, we construct controlled prompt pairs that express similar affective content using different linguistic realizations. These pairs preserve the target emotion, scene, and intended affective content while changing how that content is described. They therefore allow us to test whether semantically similar instructions lead to different generation outcomes.

Prompt Control and Validation. We apply three constraints to reduce unintended variation. First, strong–weak pairs share the same emotion, scene, base description, and template and differ only in preregistered intensity expressions; all $224/224$ pairs pass character-level structural verification. Second, cue segments are reused across templates whenever possible. Of 384 reused segments, 372 are identical and 12 differ only in pronoun disambiguation; all 384 match after normalization.

Third, controlled template pairs vary only the preregistered prompt factor, and all 448 prompts have unique SHA-256 hashes. These checks (Table B.4) establish structural consistency for controlled comparisons.

3.3 Automated Evaluation and Human Annotation Protocol

Automated evaluation and human annotation follow the same three-level decision structure and information boundaries.

Shared Evaluation Protocol. Both evaluators first determine whether the target emotion is present. If present, perceived intensity is labeled as strong or weak; otherwise it is labeled "not matched." Each of the ten affective cue dimensions is then judged independently as present or absent in support of the target emotion. Emotion and intensity are judged directly and are not inferred from cue labels.

Both evaluators see only the generated video, including its native audio, and the name and definition of the target emotion. They do not have access to the generation prompt or other metadata. The target emotion is provided because the task evaluates whether a specified emotion is expressed. The target intensity is hidden to reduce label-following bias.

Automated Evaluation. All 5,376 videos are evaluated once using Gemini 3.1 Pro with temperature 0 under a frozen two-stage protocol: the first stage judges target emotion and perceived intensity, and the second stage judges the ten cue channels (Appendix C). Videos are sampled at 8 fps. Auxiliary programs provide measurement summaries for visual dynamics and sound using cross-frame motion and audio features, respectively. The resulting single-run outputs are used for all 12-model benchmark results.

To assess evaluator stability, we conduct two additional evaluation runs on the 896-video development set. Development-set validity results are reported as the mean over the three runs (Appendix D). Wan-2.2, HunyuanVideo-1.5, Kandinsky-5.0-T2V-Pro, and LongCat-Video-13.6B do not provide usable native audio, so the sound dimension and Full 10D aggregate are not reported for these models.

Human Reference and Consensus. The human reference contains a development set for protocol calibration and a held-out test set for cross-generator validation. All first-pass annotations are collected without access to automated evaluation outputs.

The development set contains 896 videos from two generation models. Two senior annotators independently label all videos. Their matching labels form the first-pass consensus, while disagreements are resolved by the majority of available labels from five additional trained annotators (Appendix D). The test set contains 224 videos from a third generation model and was not used in protocol development. Two trained annotators provide full first-pass annotations and form the initial consensus.

After the automated evaluator is frozen, senior annotators conduct a second review of both sets (Appendix D). Reviewers see only the video and decision rules, without access to the generation prompt, model identity, target intensity, or automated outputs. Reviewed labels form the reviewed human reference, and fields without a reviewed label are excluded from analyses against it.

Metric Computation. The joint emotion-and-intensity score is computed over all 448 videos per model. Visual 9D averages cue support over the nine non-audio channels and is available for all models; Full 10D adds the sound channel and is therefore reported only for audio-capable models (Appendix C). Both are computed only for videos that successfully express the target emotion and should therefore be interpreted together with target-emotion accuracy. "Undecided" and "not applicable" labels are excluded from the corresponding denominators.

4 Results and Analysis

We analyze EmoTrace-Bench from three perspectives. First, we compare 12 T2V models in terms of target-emotion generation, intensity control, and cue realization. Second, we use controlled prompt pairs to examine the effects of cue specification, cue combination, and linguistic realization. Third, we assess the reliability of the automated evaluator through comparisons with human judgments and an alternative evaluator backbone.

4.1 Emotion Generation and Intensity Control

The 12 models exhibit distinct capability profiles. Target-emotion accuracy ranges from 49.3% to 87.0%, joint emotion-and-intensity accuracy from 29.9% to 55.1%, and conditional Visual 9D expression depth from 62.7% to 83.5%. Paired intensity-control success is substantially lower, ranging from 10.3% to 29.5%. Taken together, these results suggest that models are generally more successful at expressing the target emotion than at reliably controlling its perceived intensity.

Figure 3(a) summarizes the overall model results. Prompt-level paired bootstrap analysis of target-emotion accuracy (Appendix C) yields six bootstrap-based groups of models for which no significant within-group differences are detected under the adopted procedure. Seedance-2.5 and Veo-3.1 form the highest-performing group.

These differences are not an artifact of generation seed. Regenerating all 448 videos for Kling-3.0 under three generation seeds, with prompts and all other exposed settings fixed, moves target-emotion accuracy by 3.6 points (72.5–76.1%) and joint emotion-and-intensity accuracy by 3.8 points (45.5–49.3%), with no pairwise difference significant under exact McNemar tests (Appendix A.3). For this model and these three seed settings, the observed score ranges are approximately 7–11 times smaller than the corresponding cross-model spreads.

Performance also varies across emotion categories (Figure 3(b)). Contentment (85.9%) and amusement (85.3%) achieve the highest average target-emotion accuracy, whereas anger (52.2%) and fear (59.8%) are the lowest. Across the emotion categories included in our benchmark, the four positive emotions achieve 17.7 percentage points higher target-emotion accuracy on average than the four negative emotions, with the difference consistent in direction across all 12 models.

Intensity control exhibits a different failure pattern. Across models, 17.9%–48.2% of strong–weak pairs receive the same perceived intensity, whereas reversals occur in at most 5.8% of pairs (Table E.3, Appendix E). Thus, when intensity control fails, it is much more often because strong and weak prompts produce the same perceived intensity.

Emotion generation: overall performance and emotion-specific accuracy

Model	Audio	Tier	a Overall performance: Correct (%) Emotion	Correct (%) Emotion + intensity	Expression | emotion correct Visual 9D	Expression | emotion correct Full 10D‡	b Accuracy by target emotion: Positive emotions Amusement	Excitement	Awe	Contentment	Negative emotions Anger	Fear	Disgust	Sadness
Seedance-2.5	Yes	1	**87.0**	54.2	**83.5**	**83.4**	98.2	80.4	89.3	96.4	75.0	89.3	71.4	96.4
Veo-3.1	Yes	1	84.8	55.1	76.7	75.2	100.0	89.3	85.7	94.6	73.2	82.1	75.0	78.6
Wan-3.0	Yes	2	79.9	52.0	72.4	73.0	92.9	78.6	83.9	92.9	53.6	78.6	64.3	94.6
HappyHorse-1.1	Yes	2	79.7	53.3	77.6	77.4	94.6	75.0	80.4	89.3	71.4	69.6	67.9	89.3
MiniMax-H3-Max (fal)	Yes	3	74.3	49.1	72.0	70.0	87.5	73.2	75.0	87.5	60.7	64.3	67.9	78.6
Kling-3.0	Yes	3	73.2	44.2	74.1	70.8	80.4	73.2	73.2	89.3	62.5	66.1	67.9	73.2
MiniMax-H3	Yes	4	67.9	45.1	68.3	68.1	78.6	66.1	73.2	87.5	46.4	51.8	67.9	71.4
LTX-2.5	Yes	4	67.2	48.9	69.8	67.1	71.4	71.4	71.4	73.2	60.7	69.6	53.6	66.1
Wan-2.2	No	4	67.4	41.7	63.8	—	92.9	78.6	82.1	87.5	39.3	51.8	60.7	46.4
HunyuanVideo-1.5	No	5	55.6	38.6	62.7	—	75.0	51.8	53.6	78.6	42.9	32.1	58.9	51.8
Kandinsky-5.0-T2V-Pro	No	5†	50.5	33.7	65.3	—	66.1	57.1	42.9	66.1	28.6	46.4	41.1	55.4
LongCat-Video-13.6B	No	6	49.3	29.9	65.1	—	85.7	64.3	48.2	87.5	12.5	16.1	42.9	37.5
Mean	—	—	**69.7**	**45.5**	**71.0**	**73.1**	**85.3**	**71.6**	**71.6**	**85.9**	**52.2**	**59.8**	**61.6**	**69.9**

Score (%) · darker = higher

0 25 50 75 100

‡ Visual 9D excludes A1 and supports all-model comparison. Full 10D includes A1 and is reported only for 8 audio models; its mean uses those 8 models.
Expression scores are conditional on correct emotion. Each emotion: n ≈ 56. Single-cell CI ≈ ±13 pp.
Tier 5†: Kandinsky-5.0-T2V-Pro is sensitive to the bootstrap seed; the CI lower bound against the preceding tier is near 0.

Figure 3: Emotion generation across 12 T2V models. (a) Overall performance: target-emotion accuracy (Emotion), joint emotion-and-intensity accuracy (Emotion + intensity; E+I in the text), Visual 9D and Full 10D cue support, and bootstrap group; models within a group are not ordered. (b) Target-emotion accuracy for each of the eight emotions. Scores are obtained from one blinded automated evaluation per video.

Cue-level results reveal additional cross-model differences (Figure 4, left; Table E.1, Appendix E). Visual dynamics shows the largest variation across models, followed by event/reaction, sound, and lighting, whereas temporal organization and facial expression vary less. Because human–automated agreement is only moderate for composition and scene/space (Appendix D), we report these two channels descriptively without further cross-model interpretation.

Overall, target-emotion generation, intensity control, and cue realization capture different aspects of model behavior. These results motivate reporting these dimensions separately rather than collapsing them into a single aggregate score.

4.2 Effects of Controlled Prompt Design

We next examine how model outputs change under controlled prompt variations. The analysis focuses on three factors: explicit cue specification, cue combination, and linguistic realization. Per-model results for all 14 templates are reported in Table E.4 (Appendix E).

Explicit Cue Specification. Explicitly specifying character reactions produces the largest and most consistent gains. Adding a character reaction to a scene-only prompt improves joint emotion-and-intensity accuracy by 25–59 percentage points across models, with a mean gain of 45 points. Adding a reaction to an event prompt yields a 22–69 point gain, with a mean of 46 points. After Holm correction, these improvements are significant for 9/12 and 7/12 models, respectively. Current models therefore follow explicit facial and bodily reactions more reliably than they infer such reactions from scene or event context alone.

For cues that are not explicitly specified, we still record whether they appear in the generated video. Figure 4 compares overall cue support with support under unspecified conditions. These results characterize cue co-occurrence only and do not support causal interpretation.

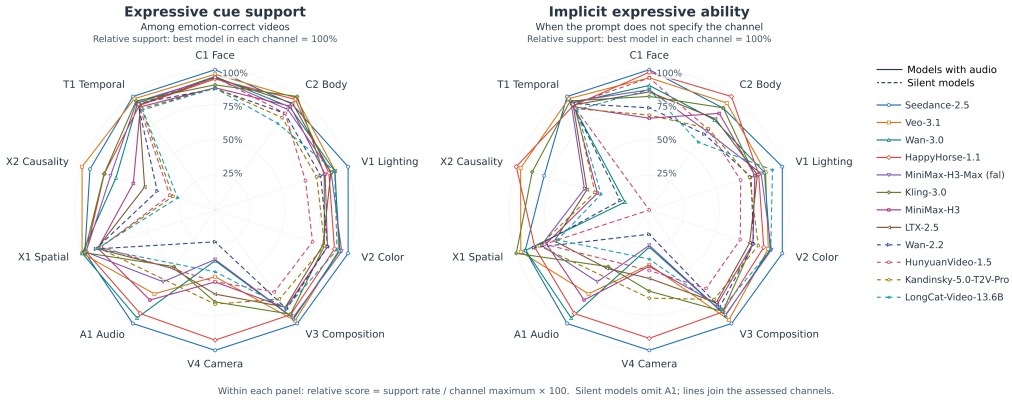


Figure 4: Overall (left) and implicit (right) cue-expression profiles of 12 T2V models. Values are relative to the best model in each channel.

Cue Combination. The benefit of adding a cue depends on which cue is added, not on how many are present. Adding a facial expression to a body-only prompt raises joint emotion-and-intensity accuracy from 40% to 64% on average (+24 points; significant for 2/12 models), whereas adding bodily behavior to a facial-expression prompt adds only 7 points (57% to 64%; significant for no model). Descriptively, facial expression contributes more than bodily behavior in this contrast. Likewise, the template that combines all explicitly writable cues (ALL_DYNAMIC, 64%) performs no better than the face-and-body prompt and only 2 points above the emotion-label baseline (EMO, 62%; Table E.4).

Alternative Linguistic Realizations. Describing a scene by how it feels conveys emotion more effectively than specifying how it should be filmed. With a character present, experiential atmosphere descriptions reach 27% joint emotion-and-intensity accuracy on average, compared with 13% for matched cinematic instructions; 9/12 models favor the experiential form, 2/12 significantly after correction. In empty scenes the two formulations perform alike (38% vs. 37%). For short, single-scene prompts of this kind, experiential language elicited emotional expression more often in our data.

All pre-specified contrasts use exact paired McNemar tests with model-wise Holm correction (Table B.3). Because each condition contains 32 videos (16 strong–weak pairs), we read these comparisons as diagnostic evidence of how models respond to specific prompt changes.

4.3 Agreement Between Automated and Human Evaluation

Because the preceding analyses rely on automated scores, we compare the automated evaluator with human judgments at both the annotation and model levels. The evaluation uses a development set of 896 videos and a test set of 224 videos from a third generation model, held out from protocol development.

Human Agreement. On the development set, pairwise human agreement is 85.7% for target emotion, 77.9% for intensity, and 80.8% for the ten cue dimensions, with Fleiss' κ values of 0.65, 0.50, and 0.60, respectively (Fleiss, 1971). On the test set, agreement between the two full-coverage annotators is 81.7%, 75.5%, and 80.2%, with Cohen's κ values of 0.54, 0.50, and 0.58, respectively (Cohen, 1960). These agreement levels characterize the reliability of the human annotations used to construct the reference labels for evaluating the automated judgments (Table D.1 in Appendix D and Table 2).

Table 2: Agreement between the automated evaluator and the human reference.

Reference	Target emotion	Conditional intensity	Joint three-way	Ten dimensions
Development, first-pass human reference	88.1 / .67	79.1 / .58	73.4 / .59	83.0 / .63
Test, first-pass human reference	84.7 / .59	88.3 / .76	77.0 / .65	80.5 / .59
Development, reviewed human reference	89.9 / .72	93.0 / .86	84.8 / .76	85.1 / .67
Test, reviewed human reference	90.2 / .76	92.0 / .84	87.1 / .81	83.0 / .64

Values are agreement (%) / Cohen's κ. Conditional intensity is computed only when both sides judge the target emotion present. Joint three-way evaluation includes "not matched," "weak," and "strong." Development-set values are means over three independent runs; full standard deviations and valid counts are reported in Table D.3. The first-pass reference is the annotator consensus before senior review; the reviewed reference is the final label set after a second review by senior annotators. Neither was produced with access to automated outputs.

Automated–Human Agreement. Table 2 reports agreement against the first-pass and the reviewed human reference. Against the first-pass reference, agreement is substantial on target emotion and the ten cue dimensions ($\kappa = 0.59$–0.67) and moderate to substantial on conditional intensity ($\kappa = 0.58$–0.76); against the reviewed reference it is 2–14 points higher, reflecting the 308 and 297 fields revised at review and, on the test set, the 41 videos that receive a label only after review (Appendix D).

Model-Level Consistency. Aggregate automated and human scores differ by at most 4 percentage points on the two development-set models and by 4–6 points on the test-set model, where automated scores are conservatively lower. Across model–emotion cells, Spearman correlation is 0.94 for target emotion and 0.90 for cue pass rates across model–emotion–dimension cells, indicating strong cell-level consistency across the three human-annotated generators and eight emotions even where absolute scores shift slightly.

Judge-Backbone Sensitivity. Applying an independent Qwen3.8-Omni-Flash evaluator to the 224 test-set videos under the same frozen protocol yields moderate to substantial agreement with Gemini across target emotion, joint intensity, the ten cue dimensions, and sound ($\kappa = 0.62$–0.68), and comparable agreement with the first-pass human reference (Appendix D.5). On this test-set generator, the evaluation patterns are therefore not tied to a single evaluator backbone; cross-evaluator replication of the full 12-model benchmark remains future work.

5 Conclusion

We introduce EmoTrace-Bench, a fine-grained diagnostic benchmark for emotion generation in text-to-video models. It evaluates target emotion, perceived intensity, and ten audiovisual cue dimensions using controlled prompts that vary intensity, cue specification, and linguistic realization. Experiments on twelve T2V models show that emotion generation, intensity control, and cue realization are distinct capabilities. Overall, EmoTrace-Bench provides a structured way to localize emotional generation failures to emotion category, intensity, and specific expressive cues, offering a more interpretable basis for evaluating and improving controllable emotional video generation.

AI USE STATEMENT

Generative AI systems played two distinct roles in this work. First, the 12 text-to-video models evaluated in the benchmark were the objects of study; Gemini 3.1 Pro was used as the automated evaluator under the protocol described in Section 3.3, and Qwen3.8-Omni-Flash (API ID: `qwen3.8-omni-flash`) was used as a second evaluator in the backbone-sensitivity analysis (Appendix D.5). Their outputs were compared against human annotations and were not treated as ground truth.

Second, generative AI tools assisted with language polishing of the authors' finalized manuscript (improving clarity, grammar, and readability) and with writing deterministic research scripts. The authors reviewed every AI-assisted passage to confirm that it preserved the intended meaning of the original text and manually reviewed all AI-assisted code. Generative AI was not used to create the benchmark prompts, formulate the conceptual framework or hypotheses, design the experiments or evaluation protocol, generate or impute human annotations, or interpret the experimental results. The authors take responsibility for the final content of the paper and all AI-assisted artifacts.

ETHICS STATEMENT

The study evaluates model-generated videos and includes human annotation of 1,120 videos, of which 896 are used for development-set validity analysis and 224 form the test set. The annotation activity underwent an institutional compliance review and was determined not to require formal review by an institutional review board or independent ethics committee; this determination should not be interpreted as a formal IRB exemption. All annotators were adults who were informed about the task and participated voluntarily. They were compensated, and they could skip content or withdraw from the task. None of the annotators, including the two senior reviewers, are authors of this paper. The submission does not include annotator-level provenance or annotator demographic information. Because the benchmark covers anger, fear, disgust, and sadness, annotators could encounter negatively valenced generated content.

EmoTrace-Bench may have dual-use implications: a diagnostic method for identifying the cues that convey emotion could also be used to make generated media more emotionally persuasive. We therefore recommend that cue-level automated scores not be used as optimization targets without human validation. The benchmark is intended for transparent evaluation and analysis of model behavior rather than for profiling individuals or inferring emotions from real people.

REPRODUCIBILITY STATEMENT

Section 3 specifies the prompt construction, structural verification, evaluation information boundaries, automated evaluation procedure, human annotation protocol, aggregation rules, and metric denominators used in this study. The submission's supplementary materials contain all 448 prompts and their per-prompt metadata, together with the annotation guidelines and three-level decision rules. Sample-level automated scores, the audit, scoring, table-generation, and plotting scripts, a subset of generated videos, the frozen evaluator prompts and rule cards, and the full video set with its manifest will be released upon acceptance. Sample-level human reference labels and annotator-level provenance are not released; the appendix reports the consensus-construction procedure and aggregate agreement statistics. These materials are provided so that the reported prompt structure and aggregate results can be independently inspected.

REFERENCES

Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229, 2012. doi: 10.1126/science.1224313.

Katalin Bálint, Thomas Klausch, and Tibor Pólya. Watching closely: Shot scale affects film viewers' theory of mind tendency but not ability. *Journal of Media Psychology*, 30(3):150–159, 2018. doi: 10.1027/1864-1105/a000189.

David Bordwell, Kristin Thompson, and Jeff Smith. *Film Art: An Introduction*. McGraw-Hill, 12th edition, 2020.

Zhe Cao, Tao Wang, Jiaming Wang, Yanghai Wang, Yuanxing Zhang, Jiahao Wang, Jialu Chen, Miao Deng, Chenxi Liao, Yize Zhang, Yubin Guo, Zhaoxiang Zhang, and Jiaheng Liu. T2AV-Compass: Towards unified evaluation for text-to-audio-video generation. *CoRR*, abs/2512.21094, 2025. doi: 10.48550/ARXIV.2512.21094. URL https://arxiv.org/abs/2512.21094.

Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics: Structured taxonomy and evaluation for professional video generation. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4aa9178b583977e31dcd11eb4fcbb38-Abstract-Conference.html.

Michel Chion. *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

Annabel J. Cohen. Music as a source of emotion in film. In *Handbook of Music and Emotion: Theory, Research, Applications*, pp. 878–908. Oxford University Press, 2010. doi: 10.1093/acprof:oso/9780199230143.003.0031.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Jeffrey F. Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. In *Handbook of Emotion Elicitation and Assessment*, pp. 203–221. Oxford University Press, 2007. doi: 10.1093/oso/9780195169157.003.0014.

Mark Coulson. Attributing Emotion to Static Body Postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2):117–139, 2004. doi: 10.1023/B:JONB.0000023655.25550.be.

Nele Dael, Marcello Mortillaro, and Klaus R. Scherer. The Body Action and Posture Coding System (BAP): Development and reliability. *Journal of Nonverbal Behavior*, 36(2):97–121, 2012a. doi: 10.1007/s10919-012-0130-0.

Nele Dael, Marcello Mortillaro, and Klaus R. Scherer. Emotion expression in body action and posture. *Emotion*, 12(5):1085–1101, 2012b.

Benjamin H. Detenber, Robert F. Simons, and Gary G. Bennett. Roll 'em!: The effects of picture motion on emotional responses. *Journal of Broadcasting & Electronic Media*, 42(1):113–127, 1998.

Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li jia Li, and Yongxin Ni. Video-bench: Human-aligned video generation benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18858–18868, 2025. doi: 10.1109/CVPR52734.2025.01757.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hua_VABench_A_Comprehensive_Benchmark_for_Audio-Video_Generation_CVPR_2026_paper.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060. URL https://doi.org/10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22139–22149. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02090. URL https://doi.org/10.1109/CVPR52733.2024.02090.

Fei Ma, Zebang Cheng, Minghui Li, Hongbo Xu, Yuyong Tan, Yihua Shao, Hanling Wang, Zhou Liu, Yuqing Gao, Dong Wang, Laizhong Cui, and Nicu Sebe. Hug-vis: A multimodal benchmark for human-centered understanding and generation in visual intelligence. *CoRR*, abs/2608.26517, 2026. doi: 10.48550/ARXIV.2608.26517.

Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005. doi: 10.3758/BF03192732.

Zongyang Qiu, Bingyuan Wang, Xingbei Chen, Yingqing He, and Zeyu Wang. Emovid: A multimodal emotion video dataset for emotion-centric video understanding and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8612–8620, 2026. doi: 10.1609/aaai.v40i10.37813.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi: 10.1037/h0077714.

Greg M. Smith. *Film Structure and the Emotion System*. Cambridge University Press, 2003. doi: 10.1017/CBO9780511497759.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 8406–8416. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00787. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Bingyuan Wang, Baistan Zhyldyzbekov, Kunyu Feng, and Zeyu Wang. Emoworld: A decoupled affective field for controllable emotional video generation. *CoRR*, abs/2608.06231, 2026a. doi: 10.48550/ARXIV.2608.06231.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, Fanshu Ding, Yushu Wang, Han Wu, Jingjing Chen, Chongxiao Wang, Yanhao Wu, Chenglong Huang, Xiaoqian Zhu, Jie Tian, Hua Li, Jingjing Fan, Mingshuang Tang, Zhong Li, Hengxia Qiang, Weibin Chen, Jinyang Zhen, Bing Zhao, Lin Qu, Jing Li, and Hu Wei. Filmbench: A film-grade benchmark for cinematic video generation. *CoRR*, abs/2607.24241, 2026b. doi: 10.48550/ARXIV.2607.24241. URL https://arxiv.org/abs/2607.24241.

Qinglan Wei, Xiaoyan Duan, Hongjiang Xiao, Ruiqi Xue, and Yuhao Wang. FG-ECVG: Fine-grained emotion-controllable video generation algorithm. *Journal of Computer-Aided Design & Computer Graphics*, 37(3):396–406, 2025. doi: 10.3724/SP.J.1089.2024-00302. URL https://www.jcad.cn/en/article/doi/10.3724/SP.J.1089.2024-00302.

Sijing Wu, Yunhao Li, Huiyu Duan, Yanwei Jiang, Yucheng Zhu, and Guangtao Zhai. HVEval: Towards unified evaluation of human-centric video generation and understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13376–13383. ACM, 2025. doi: 10.1145/3746027.3758299.

Bingyuan Zhang, Xulong Zhang, Ning Cheng, Jun Yu, Jing Xiao, and Jianzong Wang. Emotalker: Emotionally editable talking face generation via diffusion model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8276–8280. IEEE, 2024. URL https://ieeexplore.ieee.org/document/10447505.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *CoRR*, abs/2503.21755, 2025. doi: 10.48550/ARXIV.2503.21755. URL https://doi.org/10.48550/arXiv.2503.21755.

APPENDIX

A EXPERIMENTAL SETUP

A.1 MODELS AND BENCHMARK SCALE

We evaluate 12 publicly available text-to-video systems (Table A.1). Six open-weight systems were run locally from their public releases; the remaining six were accessed through hosted interfaces. MiniMax-H3 and MiniMax-H3-Max (fal) therefore differ in both the checkpoint (H3-Max is a third-party post-trained variant of H3 served by fal) and the inference channel, and the gap between them is not interpreted as a pure model difference. We identify hosted systems by the public product name available at the time of collection and open-weight systems by their public release name. Exact provider-side checkpoint hashes and API revision strings were not exposed consistently and are therefore not used to imply a finer version match than the public names support.

The benchmark was frozen in September 2026. The current release records the benchmark freeze rather than a separate access date for every provider call. This distinction matters for hosted systems, whose implementation may change without a public checkpoint identifier. Table A.1 therefore reports the public model designations and the output property that is relevant to cross-model scoring, rather than unverifiable backend revisions.

A.2 GENERATION PROTOCOL

Table A.2 summarizes the generation controls. Each system is run through a publicly accessible hosted interface or its released implementation with its recommended text-to-video configuration, and every model receives the same frozen prompt text; provider-native prompt processing, where present, is noted per model in Table A.1. The protocol therefore fixes the textual input and the number of retained samples, but does not claim that provider-specific samplers, guidance scales, or stochasticity are numerically equivalent; sensitivity to the seed setting is examined for one model in Appendix A.3.

Resolution and frame rate are left at each provider's native output and do not enter any score; the automated evaluator resamples all videos at 8 fps (Appendix C). For models without usable native audio (Table A.1), the sound channel is excluded rather than assigned a score of zero.

A.3 GENERATION-SEED SENSITIVITY

To test whether benchmark scores reflect a reproducible model-level capability rather than an idiosyncratic realization from a single generation seed, we generated three complete 448-video batches for Kling-3.0. The model version, resolution, duration, and all other exposed generation parameters were held constant, and only the generation seed was varied across the three batches (labeled A, B, and C below). The native generation pipeline was retained. Expanded prompts were not identical across batches; therefore, this experiment measures end-to-end batch variability. Each video was evaluated once under the frozen protocol; results are reported in Tables A.3–A.5. The leaderboard batch of Kling-3.0 was generated under a separate fixed seed and is not pooled into these statistics.

Across the three settings, the two binary metrics vary by less than 4 percentage points, whereas the emotion-conditioned Visual 9D and Full 10D macro scores vary by less than 0.4 points (Table A.3; the emotion-matched subsets contain 338, 325, and 341 videos). No pairwise McNemar test reaches $p < .05$ (Table A.4). A sensitivity analysis that resamples the 16 scenario families instead of prompts yields the same conclusion. Same-prompt agreement between batches (Table A.5) is only moderate for target emotion ($\kappa \leq 0.58$) and substantial for the ten dimensions ($\kappa \approx 0.62$), so aggregate stability coexists with substantial prompt-level variation; per-emotion ranges reach 14.3 percentage points on 56 videos. The analysis covers three seed settings of one model; it does not establish equivalence, characterize the full seed distribution, or separate generation variability from single-run evaluator noise.

Table A.1: **Systems included in the 5,376-video evaluation.** Each model generates one video for each of the 448 prompts. "Access" distinguishes systems reached through a hosted interface from open-weight systems run locally; "Native audio" indicates whether the generated output contains a usable model-produced audio track.

Public model designation	Access	Videos	Native audio
Seedance-2.5	hosted	448	yes
Veo-3.1	hosted	448	yes
Wan-3.0	hosted	448	yes
HappyHorse-1.1	hosted	448	yes
MiniMax-H3-Max (fal)	hosted	448	yes
Kling-3.0[†]	hosted	448	yes
MiniMax-H3	open weights, local	448	yes
LTX-2.5	open weights, local	448	yes
Wan-2.2	open weights, local	448	no
HunyuanVideo-1.5	open weights, local	448	no
Kandinsky-5.0-T2V-Pro	open weights, local	448	no
LongCat-Video-13.6B	open weights, local	448	no

[†]Native prompt processing enabled; generation seed fixed within each experiment.

Table A.2: **Generation controls and cross-model treatment.** "Native" means that a property is preserved from the provider output rather than normalized by post-processing.

Field	Protocol
Prompt submission	Identical frozen English prompt for all 12 models
Samples per prompt	One retained generation per model–prompt pair; no best-of-n selection
Duration	Approximately five seconds (provider-native duration closest to the target), continuous single-shot target; the full clip is evaluated and duration does not enter any score
Resolution	Provider-recommended native output; not forcibly matched across models
Frame rate	Provider-native output; automated evaluation samples at 8 fps
Audio	Original model-produced track retained; sound excluded when no usable native track exists
Random seed	Fixed where exposed by the provider; provider default otherwise
Post-processing	No content-based trimming, enhancement, or candidate selection before scoring

A.4 Retention and Failure Handling

The protocol retains weak, off-target, visually poor, or otherwise unsuccessful generations because these are model outcomes rather than technical failures. Only process failures—for example, an unrecoverable provider error or a file that remains undecodable after retry—are separated from generation quality. Transient network, provider, and decoding failures are retried under the same frozen request settings. A sample that remains invalid is marked `unscorable`, logged separately, and removed from the relevant denominator; it is never converted into a negative emotion label. Safety refusals and corrupted media are likewise logged as failure types rather than silently replaced by a preferred candidate.

The release manifest links every valid output to its model designation, prompt identifier, prompt hash, and media path. This one-to-one mapping supports an audit of the expected 12×448 grid without exposing model identity to either the automated evaluator or human annotators.

Table A.3: **Kling-3.0 score variation across seed settings.** All batches use the same 448 prompts and exposed generation configuration; only the generation seed is changed. Visual 9D and Full 10D are emotion-conditioned macro averages on each batch's own emotion-matched subset. SD is the descriptive standard deviation over the three settings (divisor 3) and does not estimate variance over unseen seeds.

Metric	Seed A	Seed B	Seed C	Range	SD
Target emotion (%)	75.45	72.54	76.12	3.57	1.55
Emotion + target intensity (%)	47.77	45.54	49.33	3.79	1.56
Visual 9D macro (%)	72.90	72.92	72.74	0.18	0.08
Full 10D macro (%)	69.72	70.03	69.72	0.31	0.15

Table A.4: **Prompt-matched pairwise differences between seed batches.** Δ is the right-hand seed minus the left-hand seed in percentage points; intervals are 95% percentile intervals from 5,000 paired prompt bootstrap replicates; p is from the two-sided exact McNemar test, applicable to the two binary metrics only.

Comparison	Metric	Δ	95% CI	McNemar p
A $\rightarrow$ B	Target emotion	-2.90	$[-6.92, +1.12]$	.203
	Emotion + target intensity	-2.23	$[-6.92, +2.23]$	.373
	Visual 9D macro	$+0.02$	$[-2.39, +2.53]$	–
	Full 10D macro	$+0.31$	$[-2.11, +2.68]$	–
A $\rightarrow$ C	Target emotion	$+0.67$	$[-3.12, +4.46]$	.818
	Emotion + target intensity	$+1.56$	$[-2.90, +6.25]$	.558
	Visual 9D macro	-0.16	$[-2.69, +2.30]$	–
	Full 10D macro	-0.00	$[-2.38, +2.43]$	–
B $\rightarrow$ C	Target emotion	$+3.57$	$[+0.00, +7.37]$	.076
	Emotion + target intensity	$+3.79$	$[-0.45, +8.26]$	.111
	Visual 9D macro	-0.18	$[-2.40, +2.17]$	–
	Full 10D macro	-0.31	$[-2.50, +1.86]$	–

Table A.5: **Same-prompt agreement between seed batches.** Values are agreement (%) / Cohen's κ between different generated videos for the same prompt, not repeated evaluations of one video.

Comparison	Target emotion	Three-level intensity	Ten dimensions
A vs. B	80.1 / .484	70.1 / .527	81.9 / .615
A vs. C	83.3 / .544	70.5 / .532	82.1 / .619
B vs. C	83.9 / .580	72.1 / .560	82.8 / .634

B CONTROLLED PROMPT CONSTRUCTION

B.1 FACTOR STRUCTURE

The prompt set is the complete crossing

$$8 \text{ emotions} \times 2 \text{ scenes per emotion} \times 14 \text{ templates} \times 2 \text{ intensity levels} = 448.$$

The eight emotions are amusement, excitement, awe, contentment, anger, fear, disgust, and sadness. Each emotion has two fixed scenes. Within a scene, every template is instantiated once at weak intensity and once at strong intensity, giving 224 aligned strong–weak pairs. The scenes are fixed rather than sampled, so the benchmark supports controlled comparisons over these 16 emotion–scene combinations and does not by itself establish scene-level generalization.

Each record is assembled from a base sentence and zero to five reusable cue segments. The template determines which segments are written, while the intensity member determines the preregistered expressions inside those segments. Table B.1 lists all 14 templates. "Explicit" denotes channels directly specified by the prompt; "implicit" denotes channels left unspecified but still observed in the generated video. An implicit score records cue co-occurrence and does not establish that the unspecified cue arose spontaneously or was caused by another cue.

B.2 EXPLICIT AND IMPLICIT CUE MATRIX

Table B.2 makes the prompt-side role assignment explicit. The assignment is frozen before generation and hidden during scoring. The prompt metadata covers the nine non-audio channels; temporal organization (T1) is never written directly and is observed only in the output.

This matrix also defines the limits of attribution: C1, C2, X1, and X2 each have at least one single-channel template, whereas the prompt-side effects of V1–V4 can be attributed only to the four-channel bundle (Table B.1), even though their output scores are reported separately.

Table B.1: **The 14 prompt templates.** Every template contributes 32 prompts: eight emotions, two scenes, and two intensity levels. "Words" is the median prompt length. Visual-form segments jointly specify V1–V4; these four channels are not independently manipulated. A dash means that no channel is explicit.

Template	Family	Complexity	Segments written	$\lvert\text{Exp}\rvert$	$\lvert\text{atoms}\rvert$	Words	Explicit channels	Implicit channels
EMO	emotion-label baseline	atomic	emotion label	0	1	15	–	C1, C2, V1–V4, X1, X2, T1
FACE	single channel	atomic	face	1	1	26.5	C1	C2, V1–V4, T1
BODY	single channel	atomic	body	1	1	25.5	C2	C1, V1–V4, T1
SCENE	single channel	atomic	scene appraisal	1	1	24.5	X1	C1, C2, V1–V4, T1
EVENT	single channel	atomic	event appraisal	1	1	24.5	X2	C1, C2, V1–V4, T1
FACE_BODY	paired addition	composed	face + body	2	2	42.5	C1, C2	V1–V4, T1
SCENE_REACT	paired addition	composed	scene + face + body	3	3	56	C1, C2, X1	V1–V4, T1
EVENT_REACT	paired addition	composed	event + face + body	3	3	57	C1, C2, X2	V1–V4, T1
CONTEXT_MOOD	representation	atomic	atmosphere	4	1	18	V1–V4	C1, C2, T1
CONTEXT_HARD	representation	composed	cinematic form	4	1	31	V1–V4	C1, C2, T1
FACE_TRANSITION	temporal probe	composed	facial transition	1	1	39.5	C1	C2, V1–V4, T1
ALL_DYNAMIC	stress test	stress	scene + event + face + body + atmosphere	8	5	79	C1, C2, V1–V4, X1, X2	T1
EMPTY_MOOD	no-character	composed	atmosphere	5	1	22	V1–V4, X1	–
EMPTY_HARD	no-character	composed	cinematic form	5	1	36	V1–V4, X1	–

Table B.2: **Prompt-side cue roles for all templates.** E = explicitly written; I = implicit (not written but observed); – = outside the intended scope of that template. A1 is omitted because sound is scoring-only.

Template	C1	C2	V1	V2	V3	V4	X1	X2	T1
EMO	I	I	I	I	I	I	I	I	I
FACE	E	I	I	I	I	I	–	–	I
BODY	I	E	I	I	I	I	–	–	I
SCENE	I	I	I	I	I	I	E	–	I
EVENT	I	I	I	I	I	I	–	E	I
FACE_BODY	E	E	I	I	I	I	–	–	I
SCENE_REACT	E	E	I	I	I	I	E	–	I
EVENT_REACT	E	E	I	I	I	I	–	E	I
CONTEXT_MOOD	I	I	E	E	E	E	–	–	I
CONTEXT_HARD	I	I	E	E	E	E	–	–	I
FACE_TRANSITION	E	I	I	I	I	I	–	–	I
ALL_DYNAMIC	E	E	E	E	E	E	E	E	I
EMPTY_MOOD	–	–	E	E	E	E	E	–	–
EMPTY_HARD	–	–	E	E	E	E	E	–	–

B.3 Controlled Template Pairs

In addition to the 224 within-template intensity pairs, the design contains seven cross-template contrasts, each holding emotion, scene, and target intensity fixed (Table B.3). Contrasts 1–6 form the pre-specified family analyzed with exact McNemar tests and Holm correction (Appendix C); contrast 7 is reported descriptively only (Table E.4). The replacement contrasts preserve the set of explicit channels but not text length; they are therefore descriptive comparisons of phrasing rather than isolated causal estimates of wording alone.

B.4 Representative Strong–Weak Pairs

Figure B.1 shows six prompt conditions for one fear scene as strong–weak pairs. The textual examples below are drawn from the same amusement scene. Ellipses replace only the shared base sentence; the emphasized span illustrates the intensity-bearing change.

Emotion-word baseline (EMO).

Table B.3: **Seven controlled template pairs.** Text not named in the "Intended change" column is preserved within the pair. "Tested" indicates membership in the pre-specified McNemar/Holm family.

	Pair	Operation	Intended change	Tested
1	SCENE → SCENE_REACT	add	add facial and bodily reaction to the same scene appraisal	yes
2	EVENT → EVENT_REACT	add	add facial and bodily reaction to the same event appraisal	yes
3	FACE → FACE_BODY	add	add a body segment to the same facial description	yes
4	BODY → FACE_BODY	add	add a facial segment to the same body description	yes
5	CONTEXT_MOOD → CONTEXT_HARD	replace	experiential atmosphere → explicit cinematic form, character present	yes
6	EMPTY_MOOD → EMPTY_HARD	replace	experiential atmosphere → explicit cinematic form, empty scene	yes
7	FACE → FACE_TRANSITION	replace	static facial state → facial change over time	no

> **Weak:** A woman stands beside a fairground carousel. She appears *mildly amused*.
> **Strong:** A woman stands beside a fairground carousel. She appears *intensely amused*.

Facial expression (FACE).

> **Weak:** . . . She holds a *small smile* with the lips slightly parted, both cheeks lifted, and both eyes gently narrowed.
> **Strong:** . . . She holds a *broad, open-mouthed smile*, with both cheeks raised high and both eyes narrowed.

Experiential atmosphere (CONTEXT_MOOD).

> **Weak:** . . . The space has a *light* sense of play.
> **Strong:** . . . The space has a *clear, lively* sense of play.

Cinematic form (CONTEXT_HARD).

> **Weak:** . . . The scene uses *soft bright light, slightly warm colors, a balanced medium shot, and a slow push-in*.
> **Strong:** . . . The scene uses *bright high-key light, vivid warm colors, a centered close shot, and a clear push-in*.

Figure B.2 shows frame sequences from three models for one strong-intensity prompt in each of three selected emotions; within an emotion, all models receive the same prompt.

The first two examples vary a single intensity-bearing atom. Composite templates change every preregistered intensity-bearing segment included by that template. This dose is consequently not constant across templates: the emotion-word baseline changes one atom, whereas the full stress-test template changes five. Pair-control rates are interpreted within templates rather than as evidence that templates with different doses are directly equivalent.

B.5 Structural Verification

The frozen prompt records are checked after construction rather than validated by visual inspection alone. Each record stores its scene-pair and intensity-pair identifiers, member role, explicit/implicit channel assignment, allowed intensity changes, prompt text, and SHA-256 hash. The audit results are summarized in Table B.4.



Figure B.1: Six prompt conditions for one fear scene, shown as strong–weak pairs with the intensity-bearing span highlighted. Rows 1–6 correspond to the EMO, FACE, SCENE, SCENE_REACT, CONTEXT_MOOD, and CONTEXT_HARD templates (Table B.1); rows 3–4 and rows 5–6 are contrasts 1 and 5 in Table B.3.

The 12 literal reuse exceptions (Table B.4) differ only in pronoun disambiguation and do not alter intensity or cue content. Within intensity pairs, the heuristic "the longer prompt is strong" identifies only 56.2% of pairs, so length alone does not encode the target level; length does differ across some cross-template contrasts (Table B.1, median words), which is why those contrasts are treated as descriptive (Appendix B.3).

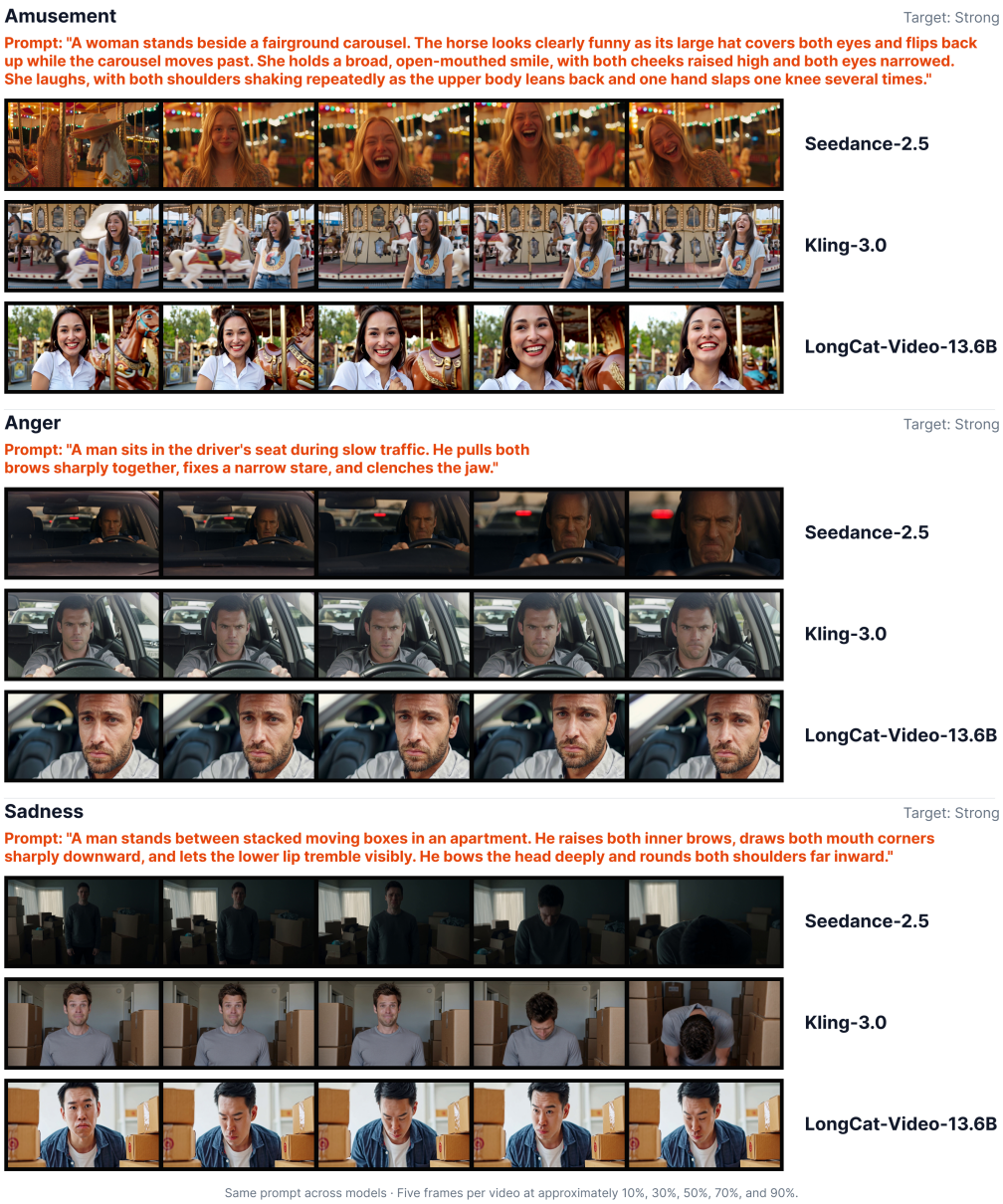


Figure B.2: Frame sequences from three models for the same strong-intensity prompt within each of three emotions. Although every video depicts the requested scene and character, the realized expression ranges from clearly strong to nearly neutral, so emotion presence and intensity are scored separately.

B.6 Released Prompt Artifact

The PDF reports the construction logic, role matrix, representative examples, and audit statistics rather than reproducing all 448 texts. The supplementary data release contains the complete prompt set and its per-record metadata, including pair pointers, channel roles, allowed changed atoms, and hashes. This keeps every reported comparison reproducible while avoiding a long verbatim prompt catalogue in the paper.

Table B.4: **Structural audit of the 448-prompt set.** Structural checks verify the frozen construction contract.

Check	Result
Total structured prompt records	448
Strong / weak members	224 / 224
Unique prompt texts	448 / 448
Unique SHA-256 prompt hashes	448 / 448
Intensity pairs passing the pairing contract	224 / 224
Literal reuse of shared segments	372 / 384
Reuse after pronoun normalization	384 / 384

C Scoring Protocol and Metrics

C.1 Three-Level Evaluation

Automated evaluation and human annotation share one ordered protocol. First, the rater decides whether the specified target emotion is present. This is a binary verification task, not free classification among the eight emotions. Second, if the target emotion is present, the rater classifies its perceived intensity as weak or strong; if it is absent, intensity is recorded as `not_matched` rather than forced. Third, the rater independently determines whether each of ten audiovisual channels contains directly observable evidence supporting the target emotion.

The three outputs answer different questions and are not converted into one another. In particular, target-emotion success and intensity are judged directly, not inferred from the number of positive cue channels. Likewise, a video may contain a local cue related to the target emotion without expressing that emotion clearly at the video level.

C.2 Operational Definition of the Ten Cue Channels

Table C.1 defines the object judged in each channel. These are diagnostic channels rather than interchangeable subscores of a single latent rating.

C.3 Shared Decision Rule

A cue channel is labeled 1 only when all five of the following conditions hold:

1. the evidence is directly observable in the video;
2. the evidence belongs to the channel being judged;
3. the evidence clearly supports the specified target emotion;
4. the evidence is not an ordinary, neutral, or incidental occurrence; and
5. the evidence can stand without borrowing evidence from another channel.

The following cases are explicitly labeled 0: no relevant evidence; evidence that is very weak or very brief; evidence that is occluded or unreadable; ambiguous meaning; a conclusion available only from the hidden prompt; a conclusion requiring an imagined backstory; evidence that depends on another channel; evidence closer to a neutral state; or evidence closer to another emotion.

The modality boundary is part of the rule. C1, C2, V1–V4, X1, X2, and T1 are judged from the picture without using the soundtrack. A1 is judged from the soundtrack without using the picture. A close-up does not by itself satisfy C1, visible movement does not by itself satisfy C2, and a reaction without a visible cause does not satisfy X2. T1 requires the target emotion itself to persist, change, or develop; camera motion, color change, or an event transition alone cannot substitute for emotional development.

Emotion-specific rule cards refine these channel-level rules. For example, a sadness–face decision requires a readable facial configuration rather than a single generic cue such as looking down, one

Table C.1: **The ten affective cue channels, grouped by source, with their operational definitions.** Each channel receives a binary judgment. A positive label requires observable evidence that specifically supports the target emotion, subject to the shared and channel-specific rules below.

Group	ID	Channel	Direct object of judgment	Channel-specific prerequisite
Character performance	C1	Facial expression	Joint configuration of eyes, brows, mouth, cheeks, and jaw	An identifiable, readable face
	C2	Bodily behavior	Posture, gesture, movement direction and energy, or meaningful stillness	An identifiable, readable body
Visual form	V1	Lighting	Brightness, contrast, shadow, direction, hardness, and their change over time	Visual content is readable
	V2	Color	Palette, warm–cool balance, saturation, color contrast, and change	Chromatic information is present
	V3	Composition	Placement, negative space, visual weight, balance, occlusion, and depth	Spatial configuration is visible
	V4	Visual dynamics	Camera motion, recomposition, stability, cuts, and meaningful holds	Sufficient temporal extent
Auditory form	A1	Sound	Voice, nonverbal vocalization, music, ambience, sound effects, and silence contrast	A usable native audio track
Context	X1	Scene / space	Openness, enclosure, scale, distance, isolation, blocking, access, and safety	Spatial relations are visible
	X2	Event / reaction	Visible trigger, affected object, reaction, and outcome	Both a trigger and its caused reaction
Temporal organization	T1	Temporal organization	Persistence, intensification, transition, or development of the target emotion	More than a momentary cue

tear, or a blue-gray palette. The same visible event can therefore receive different labels for different target emotions, but the shared five-part criterion remains fixed.

C.4 Information Boundary and Automated Evaluation

Both evaluation pathways can access the full video, its native audio track when present, and the name and definition of the target emotion. Both are blinded to the generation model, full generation prompt, requested intensity, prompt template, explicit/implicit metadata, paired sibling video, and all other ratings. Revealing the target emotion turns the first decision into target verification; hiding the target intensity prevents the strong/weak decision from copying the requested label.

Each of the 5,376 videos receives one automated evaluation pass using Gemini 3.1 Pro with temperature 0. Videos are sampled at 8 fps, and the pass consists of two separate calls under the frozen protocol above: the first call returns the target-emotion decision and, when the target is present, its perceived intensity; the second call returns the ten cue-channel judgments. No voting or cross-call aggregation is applied to the 12-model results; the two additional development-set runs (Appendix D) serve only to verify stability. Because emotion and intensity are returned together within the first call, their separation is procedural rather than an architectural claim that one output cannot influence the other; the cue judgments are produced in a separate call that does not receive the first call's output.

Two bounded auxiliary modules supply measurement summaries. The visual-dynamics module summarizes cross-frame motion for V4, and the audio module summarizes track intensity and spectral properties for A1. Neither module receives the target emotion, requested intensity, or target labels; neither emits a channel label. The automated evaluator makes the final decisions from the complete video and may not treat an auxiliary summary as a substitute for observable emotional evidence.

C.5 Metric Definitions

Let P_m be the valid videos generated by model m; the reported 12-model experiment has $|P_m| = 448$ for every model. Let $M_{mp} \in \{0, 1\}$ indicate target-emotion success, $I_{mp} \in \{\text{weak}, \text{strong}\}$ denote perceived intensity when $M_{mp} = 1$, and T_p denote the requested intensity. Let $Y_{mpd} \in \{0, 1\}$ be the score for cue channel d. Percentages below are multiplied by 100 in the reported tables.

Target emotion (Emotion).

$$\text{Emotion}(m) = \frac{\sum_{p \in P_m} M_{mp}}{|P_m|}.$$

Joint emotion and intensity (E+I).

$$\text{E+I}(m) = \frac{\sum_{p \in P_m} \mathbf{1}[M_{mp} = 1 \ \wedge \ I_{mp} = T_p]}{|P_m|}.$$

This is an unconditional conjunctive success rate: a video cannot receive credit for matching the requested degree when the target emotion is absent.

Per-channel expression depth. For a valid channel d, cue support is measured only among videos in which the target emotion succeeds:

$$Q_d(m) = \frac{\sum_{p \in P_m} M_{mp} Y_{mpd}}{\sum_{p \in P_m} M_{mp}}.$$

Thus, cue depth describes how successful emotional outputs are expressed; it does not reward a model for failing to generate the target emotion. It must be read alongside Emotion because models with different success rates contribute different conditional subsets.

Visual 9D and Full 10D. Let

$$\mathcal{D}_9 = \{\text{C1}, \text{C2}, \text{V1}, \text{V2}, \text{V3}, \text{V4}, \text{X1}, \text{X2}, \text{T1}\}, \qquad \mathcal{D}_{10} = \mathcal{D}_9 \cup \{\text{A1}\}.$$

Then

$$\text{Visual 9D}(m) = \frac{1}{9} \sum_{d \in \mathcal{D}_9} Q_d(m), \qquad \text{Full 10D}(m) = \frac{1}{10} \sum_{d \in \mathcal{D}_{10}} Q_d(m).$$

Full 10D is reported only for models with usable native audio (Table A.1); silent models receive "–", not zero, for A1 and Full 10D.

Paired intensity control. For each of the 224 aligned pairs k, let p_k^- be the weak-target member and p_k^+ the strong-target member. A pair succeeds only when both emotions are present and both perceived levels are correct:

$$\text{PairSuccess}_{mk} = \mathbf{1}\left[M_{mp_k^-} = 1 \wedge I_{mp_k^-} = \text{weak} \wedge M_{mp_k^+} = 1 \wedge I_{mp_k^+} = \text{strong}\right],$$

$$\text{PairedControl}(m) = \frac{1}{224} \sum_{k=1}^{224} \text{PairSuccess}_{mk}.$$

For failure analysis, pairs are additionally partitioned into correctly ordered, collapsed to the same perceived level, reversed, or not comparable because at least one member fails the target emotion.

C.6 Exact Denominators and Missing Values

Table C.2 collects the denominator rules used for all reported metrics. `Unscorable` denotes a process failure after retry and is never converted to a negative model outcome. A silent output is different: its visual and emotion labels remain valid, but it has no object for A1 and is excluded from audio-dependent aggregates.

For later human analyses, an intensity vote is meaningful only after that annotator marks the target emotion present. A human consensus tie is recorded as `undecided` and excluded from the corresponding agreement denominator; an unavailable channel is recorded as not applicable rather than as zero. These exclusions are reported with their valid counts so that coverage and agreement are not conflated.

Table C.2: **Exact metric denominators.**

Metric	Denominator	Treatment of missing or inapplicable values
Emotion	all valid videos (448 per model)	exclude and separately count `unscorable` process failures
E+I	all valid videos (448 per model)	emotion failure receives no joint credit; intensity is not forced
Cue channel d	videos with $M = 1$	exclude a channel only when no valid object exists, such as A1 without native audio
Visual 9D	mean of nine visual/context/temporal channel rates	always reported for a valid model run
Full 10D	mean of ten channel rates	not reported for silent models
Paired control	all 224 strong–weak pairs	a pair with either emotion absent is a failed pair, not an intensity reversal

C.7 Bootstrap, McNemar Tests, and Multiplicity

Uncertainty in target-emotion performance is estimated with 5,000 prompt-level paired bootstrap resamples. The same resampled prompt indices are applied to every model, preserving the matched benchmark design. The six reported Emotion tiers are based on 95% bootstrap intervals for paired differences. Models within a tier are not ordered, and tier assignments on a bootstrap boundary are marked as seed-sensitive.

The six pre-specified template contrasts (Table B.3, contrasts 1–6) are analyzed separately for each model. Each condition contains 32 videos. The binary outcome is joint E+I success, and the two matched conditions are compared with an exact paired McNemar test. The six planned tests within a model form one family and are adjusted with Holm's step-down procedure at $\alpha = 0.05$. Effect sizes and confidence intervals are reported alongside adjusted significance. Given the small cells, a nonsignificant contrast is treated as inconclusive rather than as evidence that the two prompt forms are equivalent.

D Human Reference and Automated-Evaluation Validity

This section reports construction of the human consensus, agreement among trained annotators, automated-evaluator agreement with the first-pass and reviewed human reference, and per-dimension performance on the test set. We report results against both the first-pass and the reviewed reference, together with their valid counts.

D.1 Human Reference and Consensus Construction

Table D.2 summarizes consensus construction. On the development set, first-pass disagreements between the two senior annotators were resolved by the majority of available labels from five trained annotators, and fields without consensus were excluded. On the test set, agreements between the two full-coverage annotators formed the first-pass consensus; disputed fields had no tie-break and were excluded from the first-pass reference. After the automated evaluator was frozen, a second review was conducted: on the development set by the same two senior annotators, on the test set by a third, senior annotator. Fields without a reviewed label are excluded from analyses against the reviewed reference. Reviewers had no access to automated outputs. The test set was generated by a generator not represented in the development set and was not used in protocol development or calibration; it was scored once under the frozen protocol.

D.2 First-Pass and Reviewed Human Reference

Relative to the development set, test-set agreement against the reviewed reference differs by $+0.3$, -1.0, and -2.1 percentage points for target emotion, conditional intensity, and the ten dimensions; video-cluster bootstrap 95% intervals for these differences are $[-4.20, +4.50]$, $[-6.47, +3.87]$, and $[-4.87, +0.47]$, all including zero. On the 183 test videos present in both references, emotion agree-

Table D.1: **Agreement among human judgments.** Development-set values summarize five trained annotators. Test-set agreement is computed only between the two full-coverage annotators; the third, senior annotator serves only as a reviewer.

Reference	Target emotion	Intensity	Ten dimensions
Development: pairwise agreement (%)	85.7	77.9	80.8
Development: Fleiss' κ	.65	.50	.60
Reviewed reference vs. trained majority (%)	90.4	90.9	94.2
Test: two-annotator agreement (%)	81.7	75.5	80.2
Test: Cohen's κ	.54	.50	.58

Table D.2: **First-pass consensus and second-review statistics.** Counts refer to fields.

Item	Development	Test
First-pass agreed fields	8,899	1,780
First-pass disputed fields	317	524
Tie-broken by trained annotators	206	–
Excluded from first-pass reference (no consensus)	111	524
Fields changed at second review	308	297
Second-review annotators	same two senior annotators	a third, senior annotator

ment rises from 84.7% to 90.7% after relabeling, and the 41 videos without a first-pass consensus reach 87.8%, so the reviewed-reference figure reflects both label revision and denominator change.

D.3 Per-Dimension Results

On the development set, seven dimensions have $\kappa \geq 0.60$ (Table D.4). Lighting, composition, and scene / space have κ between 0.52 and 0.55; each also has a human pass rate of at least 74%, indicating sensitivity to prevalence. On the test set, composition is the only dimension below $\kappa = 0.50$: a 93% human pass rate and 74% automated pass rate yield 79.3% agreement but $\kappa = 0.30$. Automated pass rates are lower than human pass rates on most test-set dimensions, whereas the two are broadly similar on the development set.

D.4 Per-Emotion Observation

Per-emotion analysis shows lower human agreement on intensity for low-arousal emotions: pairwise agreement is 70.1% for contentment and 73.4% for awe. This pattern is consistent with lower overall human agreement on intensity than on target emotion or the ten-dimensional judgments, indicating that the strong–weak boundary is intrinsically more subjective.

D.5 Judge-Backbone Sensitivity

To assess dependence on a single evaluator, we applied the frozen protocol to the 224 test-set videos with an independent Qwen3.8-Omni-Flash backbone (5 fps, temperature 0, high reasoning) and re-ran Gemini 3.1 Pro at a matched 5 fps as a control. Both evaluators used the same videos, rubric, auxiliary summaries, and two-stage structure, with identical rule-file hashes; the only Qwen-specific change was a non-semantic output-format adaptation. Qwen received no human labels, and all 224 videos and their 448 stage calls (two per video) completed without failure. Table D.5 reports agreement with the human reference and Table D.6 the direct agreement between evaluators. Against the reviewed reference, Qwen's target-emotion κ is 0.627 (95% CI $[0.508, 0.735]$) and the Qwen$-$Gemini difference is -0.095 (95% CI $[-0.212, 0.025]$), so the two are not statistically distinguishable, nor shown equivalent. Qwen's sound κ against the reviewed reference is 0.631 (81.5%), and it correctly separated 7 of 8 audio-present versus audio-removed control pairs. Because the pilot covers one generator, it does not constitute a cross-evaluator replication of the 12-model ranking.

Table D.3: **Complete agreement results between the automated evaluator and the human reference.** Values are agreement (%) / Cohen's κ; development-set values are mean $\pm$ standard deviation over three independent runs (the 12-model run plus two stability runs); test-set values are from the single run. Parentheses give valid counts.

Reference	Target emotion	Conditional intensity	Joint three-way	Ten dimensions
Development, first-pass human reference	88.1±0.7 / .670±.021	79.1±0.1 / .578±.003 (637)	73.4±0.6 / .588±.009	83.0±0.2 / .625±.003 (7,424)
Test, first-pass human reference	84.7 / .589 (183)	88.3 / .764 (94)	77.0 / .652	80.5 / .585 (1,489)
Development, reviewed human reference	89.9±0.9 / .716±.024	93.0±1.3 / .858±.026 (641)	84.8±1.8 / .764±.028	85.1±0.3 / .669±.007 (7,424)
Test, reviewed human reference	90.2 / .762 (224)	92.0 / .840 (125)	87.1 / .806	83.0 / .643 (1,639)

Column definitions follow Table 2.

Table D.4: **Per-dimension agreement against the reviewed human reference.** Each cell reports agreement (%), Cohen's κ, valid count, and the human-to-automated pass rates; run averaging follows Table D.3.

Dimension	Development	Test
Facial expression	87.9 / .74 / 768 / 64→63	84.6 / .69 / 156 / 58→52
Bodily behavior	85.1 / .70 / 768 / 53→59	84.1 / .68 / 157 / 51→52
Lighting	82.4 / .55 / 896 / 74→72	77.7 / .50 / 224 / 70→63
Color	85.3 / .64 / 896 / 71→71	83.5 / .64 / 224 / 67→63
Composition	84.9 / .55 / 896 / 80→77	79.3 / .30 / 174 / 93→74
Visual dynamics	84.0 / .64 / 896 / 64→69	86.0 / .72 / 186 / 49→48
Sound	83.3 / .66 / 896 / 40→46	85.6 / .71 / 195 / 49→52
Scene / space	84.0 / .52 / 384 / 78→80	87.5 / .65 / 96 / 80→74
Event / reaction	81.9 / .63 / 256 / 45→42	79.7 / .58 / 64 / 45→28
Temporal organization	90.8 / .76 / 768 / 74→74	82.8 / .62 / 163 / 72→62

Table D.5: **Test-set agreement of two evaluator backbones with the human reference.** Values are agreement (%) / Cohen's κ; both evaluators run at 5 fps under the frozen protocol.

Reference / task	Qwen3.8-Omni-Flash	Gemini 3.1 Pro
First-pass: target emotion	81.4 / .509	83.6 / .574
First-pass: joint three-way	70.3 / .542	73.6 / .602
First-pass: conditional intensity	84.3 / .679	82.8 / .653
First-pass: ten dimensions	83.1 / .627	79.9 / .575
Reviewed: target emotion	84.4 / .627	88.4 / .721
Reviewed: joint three-way	74.2 / .610	80.4 / .705
Reviewed: conditional intensity	81.2 / .629	85.1 / .702
Reviewed: ten dimensions	82.2 / .611	81.3 / .610

Table D.6: **Direct agreement between evaluator runs on the same test-set videos.** Values are Cohen's κ. The last row compares the 8 fps setting used in the main results with the 5 fps control.

Pair	Emotion	Joint three-way	Ten dimensions	Visual 9D	Sound
Qwen3.8-Omni-Flash vs. Gemini (5 fps)	.640	.622	.678	.676	.678
Qwen3.8-Omni-Flash vs. Gemini (8 fps)	.679	.630	.680	.682	.660
Gemini 8 fps vs. 5 fps	.875	.844	.861	–	.893

E ADDITIONAL RESULTS

This appendix reports the numeric values behind the cue-level figures in the main text and two per-model breakdowns that the main text summarizes only in aggregate: paired intensity control and joint emotion-and-intensity accuracy for each of the 14 prompt templates. All values come from the same single blinded automated evaluation per video as the main results. Rows follow the ordering of Figure 3: audio-capable models first, silent models second, and within each block by target-emotion accuracy.

E.1 CUE-LEVEL EXPRESSION PROFILES

Table E.1 gives the values plotted in the left panel of Figure 4.

Table E.1: **Cue-level expression profiles of 12 T2V models.** Each score is the percentage of target-emotion-successful videos in which the corresponding channel supports the target emotion. N is the number of such videos for each model.

		Character		*Visual form*				*Audio*	*Context*		*Temporal*
Model	N	C1	C2	V1	V2	V3	V4	A1	X1	X2	T1
Seedance-2.5	390	**86.4**	70.1	**85.1**	**84.9**	**89.7**	**93.6**	**82.6**	89.8	58.6	**93.2**
Veo-3.1	380	83.6	74.1	76.8	79.0	87.4	44.5	61.0	**91.7**	**62.3**	91.4
Wan-3.0	358	82.2	68.1	76.5	79.6	85.8	33.8	78.5	91.6	46.4	87.8
HappyHorse-1.1	357	80.7	73.1	72.8	76.2	82.6	86.8	75.3	88.8	51.9	85.9
MiniMax-H3-Max (fal)	333	81.4	66.4	74.2	80.8	85.0	33.0	52.2	89.0	49.0	88.9
Kling-3.0	328	77.1	**75.0**	69.8	69.8	82.3	61.3	41.2	90.1	52.0	89.6
MiniMax-H3	304	75.3	68.6	70.4	71.7	77.3	48.0	65.5	80.8	38.3	84.7
LTX-2.5	301	81.6	70.6	74.1	68.8	76.7	56.1	42.5	81.4	33.0	86.3
Wan-2.2	302	74.7	63.8	67.5	71.8	78.2	21.2	–	82.3	27.1	87.5
HunyuanVideo-1.5	249	81.7	63.9	57.8	62.2	64.7	47.8	–	78.5	21.2	86.8
Kandinsky-5.0-T2V-Pro	226	75.2	61.1	65.0	69.9	70.3	62.8	–	80.2	19.7	83.3
LongCat-Video-13.6B	221	75.0	57.2	76.5	77.8	77.4	41.2	–	81.1	17.5	82.2
Mean		*79.6*	*67.7*	*72.2*	*74.4*	*79.8*	*52.5*	*62.4*	*85.4*	*39.7*	*87.3*
Range		*11.7*	*17.8*	*27.3*	*22.6*	*25.1*	*72.4*	*41.4*	*13.3*	*44.8*	*11.0*

Channel abbreviations follow Table C.1. A1 is unavailable for the four silent models (–); its mean and range are computed over the eight audio-capable models. Bold denotes the largest observed value in each column and does not imply statistical significance. All ten channels have a cross-model range of at least 10 percentage points, but T1 is saturated (at least 80% for every model) and the C1 and T1 ranges only just exceed 10 points, so differences within these two columns are not interpreted.

E.2 IMPLICIT CUE EXPRESSION

Table E.2 gives the values plotted in the right panel of Figure 4. The bottom rows contrast the 12-model mean support under unspecified and specified conditions; the difference indicates how strongly each channel depends on being written in the prompt. These are co-occurrence rates and do not establish that an unspecified cue was caused by another cue.

E.3 PAIRED INTENSITY CONTROL

Table E.3 decomposes the 224 aligned strong–weak prompt pairs per model into four mutually exclusive outcomes. A pair is *correctly ordered* when the weak member is judged weak and the strong member strong; *collapsed* when both members receive the same perceived intensity; *reversed* when the strong member is judged weaker than the weak member; and *not comparable* when at least one member fails the target emotion. The dominant failure for every model is collapse rather than reversal.

Table E.2: **Implicit cue-expression profiles of 12 T2V models.** Each score is the percentage of target-emotion-successful videos, among those whose prompt does not explicitly specify the channel, in which that channel supports the target emotion. Implicit 9D is the mean over the nine visual, contextual, and temporal channels.

	Character		*Visual form*				*Audio*	*Context*		*Temporal*	
Model	C1	C2	V1	V2	V3	V4	A1[¶]	X1[§]	X2[§]	T1[¶]	Implicit 9D
Seedance-2.5	**71.5**	54.9	**81.2**	**81.6**	**88.1**	**94.6**	**82.6**	90.3	29.0	**93.2**	76.1
Veo-3.1	67.4	57.6	71.0	73.0	85.3	36.9	61.0	93.5	35.5	91.4	68.0
Wan-3.0	63.5	48.5	70.3	74.2	81.8	25.0	78.5	90.0	6.7	87.8	60.9
HappyHorse-1.1	70.2	**61.1**	66.8	70.2	79.0	86.5	75.3	83.3	36.7	85.9	71.1
MiniMax-H3-Max (fal)	61.3	42.5	68.3	74.9	81.5	23.8	52.2	78.6	17.9	88.9	59.7
Kling-3.0	57.9	55.0	64.4	63.6	78.7	54.7	41.2	96.8	32.3	89.6	65.9
MiniMax-H3	46.8	52.0	64.5	63.5	73.0	38.0	65.5	75.0	14.3	84.7	56.9
LTX-2.5	60.2	49.1	66.0	61.6	71.8	46.1	42.5	69.0	17.2	86.3	58.6
Wan-2.2	52.1	40.9	62.1	64.0	75.4	16.3	–	84.0	8.0	87.5	54.5
HunyuanVideo-1.5	67.3	43.6	55.8	55.8	60.9	40.8	–	76.2	0.0	86.8	54.1
Kandinsky-5.0-T2V-Pro	48.4	43.9	61.2	66.9	69.4	59.4	–	80.0	15.0	83.3	58.6
LongCat-Video-13.6B	61.3	36.5	75.2	74.5	77.2	33.1	–	66.7	13.3	82.2	57.8
Mean, channel unspecified	*60.7*	*48.8*	*67.2*	*68.6*	*76.8*	*46.3*	*62.4*	*81.9*	*18.8*	*87.3*	*61.8*
Mean, channel specified	*90.8*	*86.3*	*82.6*	*86.3*	*85.9*	*65.5*	–	*86.2*	*47.8*	–	
Specified − unspecified (pp)	*+30.1*	*+37.5*	*+15.3*	*+17.7*	*+9.1*	*+19.2*	–	*+4.2*	*+29.0*	–	

Channel abbreviations follow Table C.1; pp = percentage points. [¶]A1 and T1 are never specified in prompts, so all of their support is implicit and no "specified" counterpart exists; A1 is unavailable for the four silent models and its mean is computed over the eight audio-capable models. [§]X1 and X2 are unspecified in only about 30 videos per model (confidence intervals of roughly ± 15 percentage points), so these two columns are not bolded. Bold denotes the largest observed value in each remaining column and does not imply statistical significance. Channels whose cross-model range is below 10 percentage points are not interpreted.

Table E.3: **Outcomes of the 224 strong–weak prompt pairs per model.** All values are percentages of pairs and sum to 100 within a row.

Model	Correctly ordered	Collapsed	Reversed	Not comparable
Seedance-2.5	27.7	47.3	5.8	19.2
Veo-3.1	28.6	48.2	1.3	21.9
Wan-3.0	26.8	43.8	0.9	28.6
HappyHorse-1.1	**29.5**	36.6	3.1	30.8
MiniMax-H3-Max (fal)	23.2	36.6	1.3	38.8
Kling-3.0	19.6	39.7	1.3	39.3
MiniMax-H3	19.6	32.1	1.8	46.4
LTX-2.5	27.2	25.9	1.3	45.5
Wan-2.2	17.9	36.6	1.3	44.2
HunyuanVideo-1.5	21.9	20.5	0.9	56.7
Kandinsky-5.0-T2V-Pro	14.3	17.9	0.9	67.0
LongCat-Video-13.6B	10.3	22.3	1.3	66.1
Mean	*22.2*	*34.0*	*1.8*	*42.0*

Differences between adjacent models in the first column lie within sampling uncertainty (roughly ± 6 percentage points), so this table does not rank models on intensity control; bold marks the largest observed value only. The "not comparable" share follows target-emotion accuracy: the lower a model's Emotion score, the fewer pairs enter the intensity comparison.

E.4 PER-TEMPLATE JOINT ACCURACY

Table E.4 reports joint emotion-and-intensity accuracy (E+I) for each of the 14 prompt templates and each model, together with the paired differences for the six pre-specified contrasts (Table B.3); the two "experiential → cinematic" Δ rows are the character-present and empty-scene instances of the same replacement. The FACE_TRANSITION row is included for completeness; contrast 7 was not tested, and no significance is claimed for its difference from FACE.

Table E.4: **Joint emotion-and-intensity accuracy (%) by prompt template and model.** Regular rows report E+I over the 32 videos of each template. Italic Δ rows report the paired difference (in percentage points) of the row immediately above from the named template, for the six pre-specified contrasts; * denotes $p < 0.05$ under an exact paired McNemar test after model-wise Holm correction over the six contrasts. "Mean" is the 12-model mean; "Sig." counts models with a significant difference.

Template	Seedance-2.5	Veo-3.1	Wan-3.0	HappyHorse-1.1	MiniMax-H3-Max (fal)	Kling-3.0	MiniMax-H3	LTX-2.5	Wan-2.2	HunyuanVideo-1.5	Kandinsky-5.0-T2V-Pro	LongCat-Video-13.6B	Mean	Sig.
Emotion-label baseline														
EMO	66	72	81	75	72	62	62	69	62	50	44	25	62	
Character performance														
FACE	69	81	50	62	69	66	56	53	47	53	34	41	57	
BODY	56	41	47	53	38	47	25	50	28	28	38	25	40	
FACE_BODY	69	69	66	81	62	56	66	78	50	62	66	41	64	
Δ vs. FACE	*+0*	*−12*	*+16*	*+19*	*−6*	*−9*	*+9*	*+25*	*+3*	*+9*	*+31*	*+0*	*+7*	*0/12*
Δ vs. BODY	*+12*	*+28*	*+19*	*+28*	*+25*	*+9*	*+41**	*+28*	*+22*	*+34**	*+28*	*+16*	*+24*	*2/12*
FACE_TRANSITION	72	81	69	66	66	75	56	56	69	69	31	34	62	
Context with and without a character reaction														
SCENE	34	41	25	34	16	19	19	19	19	6	9	12	21	
SCENE_REACT	66	66	75	66	75	66	69	78	62	59	59	47	66	
Δ vs. SCENE	*+31*	*+25*	*+50**	*+31*	*+59**	*+47**	*+50**	*+59**	*+44**	*+53**	*+50**	*+34**	*+45*	*9/12*
EVENT	34	31	31	41	31	16	16	12	16	3	3	16	21	
EVENT_REACT	66	69	84	66	72	62	81	72	62	72	59	38	67	
Δ vs. EVENT	*+31*	*+38*	*+53**	*+25*	*+41*	*+47**	*+66**	*+59**	*+47**	*+69**	*+56**	*+22*	*+46*	*7/12*
ALL_DYNAMIC	69	66	72	69	62	53	62	78	59	75	53	50	64	
Atmosphere, character present: experiential vs. cinematic														
CONTEXT_MOOD	47	41	31	41	28	22	22	28	22	6	25	9	27	
CONTEXT_HARD	28	19	12	9	12	6	22	9	19	6	0	12	13	
Δ vs. CONTEXT_MOOD	*−19*	*−22*	*−19*	*−31**	*−16*	*−16*	*+0*	*−19*	*−3*	*+0*	*−25**	*+3*	*−14*	*2/12*
Atmosphere, empty scene: experiential vs. cinematic														
EMPTY_MOOD	34	47	44	47	44	41	38	41	31	25	25	38	38	
EMPTY_HARD	50	50	41	38	41	28	38	41	38	25	25	31	37	
Δ vs. EMPTY_MOOD	*+16*	*+3*	*−3*	*−9*	*−3*	*−12*	*+0*	*+0*	*+6*	*+0*	*+0*	*−6*	*−1*	*0/12*

Template definitions are given in Table B.1. Each cell has $N = 32$ and a confidence interval of roughly ± 17 percentage points, so nonsignificant differences are inconclusive rather than evidence of no effect. Δ values are computed from the underlying counts and may differ by one point from the difference of the rounded percentages shown. The two Δ rows under FACE_BODY share the same target template and are not independent tests. Target-emotion accuracy alone is not reported here because it saturates: four templates reach a 12-model mean of at least 85%, leaving little room to separate them.