

Verifiable Creative Direction: Parsing Free-Text Intent into Bounded Structural Deltas for Screenplay Generation

Mihai-Cristian Strejaru

National University of Science and Technology
Politehnica Bucharest
Bucharest., Romania
mihai.strejaru@stud.acs.upb.ro

Ștefan Trăușan-Matu

National University of Science and Technology
Politehnica Bucharest
Research Institute for Artificial Intelligence
“Mihai Drăgănescu”
Academy of Romanian Scientists
Bucharest, Romania
stefan.trausan@upb.ro

ABSTRACT

A screenwriter co-writing a scene with a language-model tool can ask for a change such as “make it more tense,” but the tool has no way to show whether the scene it returns is actually more tense than before. Because a request like this is represented only in words, never as a measurable quantity, the result cannot be checked afterward. We describe a system that treats such a request as a change to a small set of structural numbers instead. From a corpus of professionally written screenplays, the system learns what a well-formed scene tends to look like at a given point in a story, a target expressed as six measures: tension, mood, volatility, pacing, dialogue-to-action ratio, and peak intensity. A free-text direction is read into small, bounded changes over that target, and the same analyzer that produced the target checks the finished scene against it, closing the loop between asking for a change and verifying it happened. Most of these measures follow a plain-language direction well above chance, and reading a direction leaves the generated scene slightly farther from its structural target than leaving it undirected, a cost that persists across the measures we track.

AUTHOR KEYWORDS

controllable text generation; screenplay generation; computational narrative; human-AI co-writing; large language models; verifiable generation; narrative structure

DOI: 10.37789/icusi.2026.20

INTRODUCTION

A screenwriter who asks a co-writing tool to make a scene “more tense” is asking for a change in structure, not just in wording. Current language-model tools can accept that request as a prompt, but they work entirely in words, with no numeric goal for the scene. So after the scene comes back, there is no way to check whether the text moved in the requested direction, by how much, or whether it drifted away from the shape the story had established.

This gap has a concrete cause. A language model has no internal notion of where a scene sits inside a film. A scene at 15% of the story has a different dramatic job than one at 80%, yet an instruction such as “write a tense scene” carries none of that information. The model may well produce something tense, but whether it is tense in the right way, at the right point in the arc, is left to chance. The same problem affects a follow-up request: if the tool cannot state, as a number, what

“more tense” should mean for this scene, it also cannot report back whether it succeeded.

Two lines of prior work frame our approach. On one side, co-writing systems plan before they write. Dramatron [7] builds a logline, then characters, then a scene-by-scene synopsis, then dialogue, passing summaries down each level to hold the story together. Nothing in that plan is a measurable quantity, so there is no tension or pacing figure to check the finished dialogue against. Re³ [15] is closer to a control loop, since it drafts, scores, and revises, but its score measures coherence with earlier text, not position on a narrative arc. On the other side, quantitative work on story structure suggests that structure can be measured rather than assumed. Reagan et al. [12] reduced the emotional arcs of 1,327 stories to a small number of recurring shapes. Neither line, though, lets a screenwriter issue a plain-language direction and then see, as a number, whether the generated scene followed it.

This paper describes a system built around that missing check, and it focuses on one step in particular. Our system has two parts. The first part learns structure from data. Six structural measures are computed for every scene in a corpus of 206 professionally written movie screenplays (about 21,000 scenes), and the scenes are grouped so that, for a story of a given shape, the system can say what kind of scene is expected at a given position. This yields a numeric target for any scene we want to generate. The second part, the control loop studied in this paper, adjusts that target from a plain-language direction. A single temperature-zero call to a general-purpose language model (Llama-3.3-70B-Instruct) is given the current target and the direction, and returns a numeric change for each measure the direction affects: asking to “make it more tense,” for instance, typically returns a positive change to the tension and volatility numbers. Each change is clipped to a valid range and added to the learned target rather than replacing it. The generated scene is then measured by the same analyzer that produced the target, so the loop closes on itself.

We had already evaluated the whole system against a plain baseline, and guided generation lands closer to its structural target. That comparison, however, is dominated by the target-learning part, because a scene generated with no user direction at all still benefits from it. Whether the direction-reading step does anything measurable on its own had not been tested. This paper adds that test. Holding the target-learning and generation machinery fixed, we compare scenes generated with and without a parsed direction, and we ask two questions: does the measured change agree in direction with the requested

change, and does adding a direction cost anything in how closely the scene matches its target?

Our contributions are three. First, we describe a control loop that turns an open-ended creative direction into a numeric change that can be verified after generation, a check that planning-based and revise-and-score co-writing systems do not provide. Second, we run an ablation that isolates the direction-reading step from the target-learning part that dominated our earlier evaluation. Third, we report plainly where this directional control works and where it does not.

The paper continues with a review of related work on planning-based co-writing and quantitative story structure, then describes how the structural target is derived from the corpus, the control loop that reads a direction into bounded changes over that target, and the evaluation that isolates the direction-reading step, before turning to limitations and conclusions.

STATE OF THE ART

Planning and revision in LLM story generation. Dramatron [7] expands a short prompt into a full script through a fixed hierarchy of language-model calls, from title and characters down to a scene-by-scene outline and finally the dialogue, with each stage conditioned on generated summaries of the ones above it. Its evaluation, interviews with 15 industry professionals, gave useful qualitative feedback but no repeatable measurement, and nothing in its plan is numeric, so one cannot ask afterward whether a scene reached a particular tension level. Re³ [15] goes further toward iteration, organizing generation into plan, draft, rewrite, and edit stages in which a scoring model marks passages to revise for coherence and relevance. The check it performs, though, is coherence with what came before, not position on a structural arc. A scene can fit its neighbours and still sit far from where the story's shape would place it. Plan-and-Write [16] plans through keyword sequences instead of a prose synopsis, a richer skeleton than Dramatron's, but the skeleton is still ordinary language, written by the same model that will later expand it, so it inherits that model's structural blind spots.

Structure learned from data, not assumed. Screenwriting manuals [2, 13] lay out fixed templates with dramatic beats at set percentages of the runtime. Reagan et al. [12] took the opposite path, reducing the sentiment curves of 1,327 stories to a few recurring shapes and showing that structure can be recovered from data rather than assumed in advance. Closer to our own domain, Papalampidi et al. identify turning points directly in movie screenplays [10] and use the resulting structure to guide extractive summarization [11], treating structure as something to recover from a finished script rather than a target to write toward. Emotional arcs have also been used to guide generation directly, not only to detect it afterward: Wen et al. [14] condition procedural game-level generation on an arc shape, closer in spirit to what we do, though for level layouts rather than prose, and without a step that measures the finished output against the arc it was meant to follow. Our system extends the data-driven line to screenplay generation with that check attached. It builds a signature for each film from four of its structural measures sampled across the story, groups those signatures into a handful of arc shapes, and combines them with per-scene groupings to produce a

position-specific numeric target, rather than a single label for the whole film.

Controllable text generation. A related line of work steers a model toward target attributes rather than target structure. CTRL [4] conditions generation on control codes for properties such as topic and style; PPLM [1] nudges a frozen model's activations toward an attribute classifier during decoding. More recent work aims closer at the precision problem we care about, reaching a specific attribute value rather than a rough direction: Zhang et al. [17] treat attribute-intensity control as a target-reaching problem and use a learned value function to steer decoding toward it, a check built into the generation process itself. A recent survey [5] covers this broader space of attribute-steering methods for large language models. All of these methods, including ours, agree that a request should be readable as a number. They differ on when that number is checked. Attribute-steering methods verify against an internal classifier while decoding is still under way; our loop instead reads the direction before generation starts and re-measures the finished scene afterward, using the same external analyzer that set the target, independent of what happened during decoding. None of them carry a notion of where a passage sits on a narrative arc, and none checks, after generation, whether a requested change of a given size took place in a longer, structured piece of writing such as a screenplay scene.

Evaluating open-ended generation. Overlap measures such as BLEU and ROUGE, and embedding measures such as BERTScore [18], all need a reference text and agree poorly with human judgment on creative writing [8]. A comprehensive evaluation of large language models on a shared creative-writing task found the same difficulty at larger scale: even careful human and judge-based scoring across many models struggled to agree on which stories were better, and the ranking shifted with the criteria used [3]. This is expected, since many different scenes can satisfy the same instruction equally well, and it is part of why we lean on a reference-free structural measure instead. The structural distance we use needs no reference. The same analyzer that produced the corpus targets is run on the new scene, and its distance from the target becomes a repeatable measure of how closely the scene followed the constraint. This is a measure of compliance, not of craft. Its usefulness rests on our earlier finding that closer compliance tends to go with writing that independent judges rate more highly. Language-model judges bring their own known biases, toward earlier or longer answers and toward their own style [19]. Because the evidence in this paper is measured rather than judged, those biases are not a central concern here.

None of the systems above can say whether a specific, user-issued direction moved the generated text in the requested direction by a measurable amount. That is the gap this paper addresses.

DERIVING THE STRUCTURAL TARGET

The control loop we study adjusts a target that the system learns from data rather than one we design by hand. Six measures are computed for every scene in the corpus, and the computation is deterministic, with no learned model in the loop: tension, mood, volatility, pacing, dialogue-to-action ratio, and peak intensity (see Table 1). Tension combines three

Measure	What it captures
Tension	Dramatic conflict: polarity, swing, and confrontation combined (Eq. 1)
Mood	Aggregate scene sentiment, from dark (-1) to bright (+1)
Volatility	Spread of sentiment across the lines: how far the mood swings
Pacing	Estimated tempo, in words per second
Dialogue/action ratio	Dialogue lines divided by action lines
Peak intensity	The strongest single burst of physical action

Table 1: The six structural measures, all computed deterministically; tension follows Equation 1.

signals, polarity, swing, and open conflict,

$$T = 0.4 |\bar{s}| + 0.3 \sigma_s + 0.3 r_{\text{confr}}, \quad (1)$$

where $\bar{s}$ is the mean sentiment of the lines in the scene, σ_s is the spread of those sentiments, and r_{confr} is the share of lines that read as confrontation. Pacing comes from a simple reading-time model, with dialogue read at a conversational pace and each action line given a fixed on-screen duration. The resulting words-per-second figure is not meant to be exact, but it is applied the same way across the whole corpus, so the relative order of scenes is what matters.

The measures are turned into reusable structure by unsupervised clustering. Scene vectors are grouped with k -means into six recurring scene types, after a quantile transform that maps every measure onto a common, roughly uniform scale before clustering, so the heavy-tailed dialogue-to-action ratio, which can run from 0 to past 100, cannot dominate the distance calculation the way it would under ordinary standardization. The grouping is deliberately soft. A silhouette score of 0.33 reflects that dramatic regimes lie on a continuum rather than in cleanly separated clusters, so a label reads as “the regime this scene mostly belongs to” rather than a hard class. Reaching even this score required a length-dependent correction to one of the six measures before clustering, described in the harder-measures analysis.

At the level of a whole film, a signature is built from four of the six measures, tension, mood, volatility, and pacing, each sampled on a common 100-step timeline and standardized channel by channel, giving a 400-number curve per film. The dialogue-to-action ratio and peak intensity are left out here: the first is too skewed to compare across films, and the second largely tracks tension. These signatures cluster into five canonical arc shapes, for example a steady rise in tension or an early mid-film crisis. Crossing arc shape with ten position bins gives the lookup table. For a scene at, say, 40% of a story whose crisis falls mid-way, the table names the scene type expected there, and the center of that type supplies the six target numbers. That vector is the baseline target the control loop adjusts (Figure 1 summarizes this derivation).

THE CONTROL LOOP

Figure 2 shows the loop end to end. The rest of this section follows it left to right: reading a direction into deltas, clamping and injecting them, and verifying the result.

From direction to delta

Given the target of the target-derivation stage and a user’s free-text direction, such as “make it more tense and let the confrontation escalate,” a temperature-zero call to Llama-3.3-70B-Instruct turns the direction into numeric changes, one for

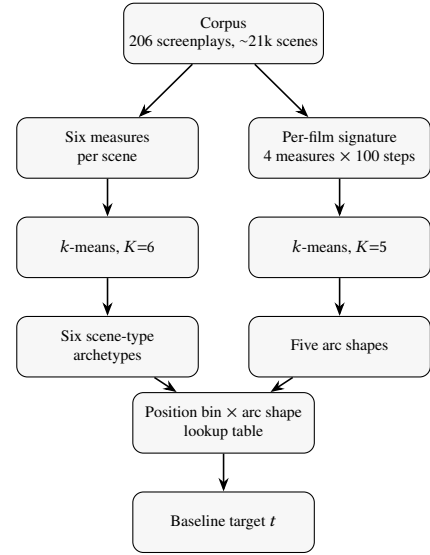


Figure 1: Deriving the baseline target: per-scene clusters (left) crossed with per-film arc shapes (right).

each measure the direction should affect. The prompt states the current target, gives the direction, and shows a single worked example:

You are a narrative engine. The current scene mathematically requires these baseline features: {...current values...} The user has provided the following direction: "<intent>" Output a JSON object containing mathematical deltas to apply. For example, for a "quiet, calm dialogue" you might output: {"tension_score": -0.3, "pacing_words_per_sec": -0.5, "dialogue_action_ratio": +2.0} Only include features that should change. Output ONLY valid JSON.

Two choices matter here. The worked example fixes both the format of the answer and the size of the numbers. Without it, the model tends to return values that look like new absolute targets instead of small changes. Running the step at temperature zero keeps the reading close to deterministic, so the same direction yields the same changes on repeated calls. That stability matters for the ablation in the validation, where the requested change needs to be a fixed quantity rather than a fresh draw each time.

Clamping and injection

The returned numbers are not applied as they arrive. Each measure has a valid range, with tension and volatility in $[0, 1]$, mood in $[-1, 1]$, and pacing in $[0.1, 2.3]$, and every change is added to the target and then clipped to that range. Pacing’s upper bound follows directly from how it is computed (the target-derivation stage): with a fixed time cost for each action line, the ratio of words to elapsed time cannot exceed 2.5 for any scene that contains action content at all, and only approaches that limit in the unrealistic case of a scene with no action lines whatsoever. An earlier version of this range allowed values up to 5, which meant a request could ask for a pacing figure no scene could ever reach; 2.3 keeps every target inside what the formula can actually produce. The change

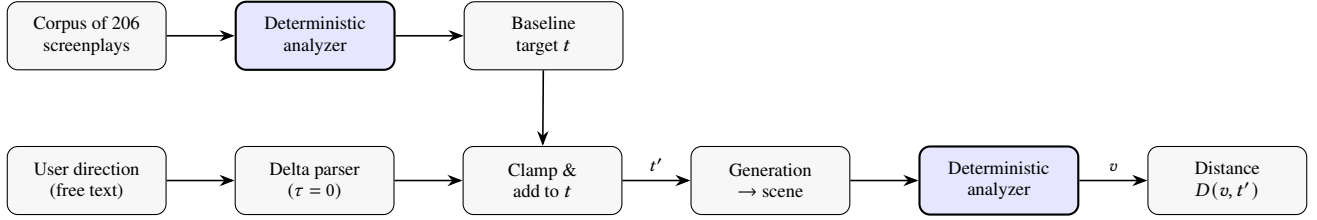


Figure 2: The control loop. The same deterministic analyzer (blue) sets the target t from the corpus and later measures the generated scene; a direction becomes clamped deltas on t , scored by the distance $D(v, t')$.

Measure	Baseline t	Adjusted t'	Change
Tension	0.12	0.37	+0.25
Volatility	0.00	0.30	+0.30
Pacing (words/s)	2.10	2.30	+0.20 (capped)
Peak intensity	0.55	0.75	+0.20
Mood	-0.10	-0.10	—
Dialogue/action ratio	0.00	0.00	—

Table 2: “Make it more tense” read against one scene’s baseline t ; dashes mark the measures it leaves untouched.

adjusts the learned target; it does not replace it. A direction that would push tension far too high for a scene’s position still lands inside the valid range, anchored to where the corpus says the scene should sit. In short, the user’s intent bends the target rather than overriding the learned structure.

Table 2 shows the step on a real scene. The direction “make it more tense and let the confrontation escalate” raises tension, volatility, pacing, and peak intensity, and leaves mood and the dialogue-to-action ratio where the corpus put them. Pacing also shows the clamp in the control loop at work: the requested rise would otherwise push the target past what the formula can produce, so the adjusted value sits at the 2.3 ceiling instead. This is the behaviour the design intends: the direction bends a few relevant measures, the clamp keeps every one of them reachable, and the rest stand as the corpus left them.

The six adjusted numbers, together with the labels of the arc shape and scene type behind them, are placed in the generation prompt as a short block, each value annotated with a one-line reading so that the writing model need not guess what a figure like “tension: 0.71” means. The block is placed near the end of the prompt rather than in the middle, since instruction-following degrades for text buried in the middle of a long context [6]. Numeric constraints cannot be reconstructed from the surrounding story context the way a scene summary can, so they are given a position where they are more reliably read.

Verifying compliance

After the scene is generated, the same analyzer that produced the corpus measures, not a separate model and not a language-model judge, is run on the new text and returns its six-measure vector v . The distance from a target t is the sum of the absolute differences across the six measures, on their natural scales:

$$D(v, t) = \sum_{j=1}^6 |f_j(v) - t_j|. \quad (2)$$

Because the measures enter this sum on their own scales, the dialogue-to-action ratio, which can run well beyond $[0, 1]$, weighs more heavily than the bounded measures. This does not distort any given comparison, since the same formula is applied to every condition, but it does mean D should be read

as an ordering signal rather than a quantity with physical units. This distance, not a judge, is what the ablation uses to check whether reading a direction costs anything in how closely the scene matches its target.

VALIDATION

Full-system comparison

We first evaluated the whole system against a plain baseline. Both conditions used the same language model for writing, Llama-3.3-70B-Instruct, and the same story context. The baseline received the story material and the scene’s surrounding context but none of the structural additions: no numeric target, no retrieval of similar scene patterns as examples, and no rule that pushes a later variant away from an earlier one. Guided generation landed markedly closer to its structural target than the baseline. A component-by-component check, however, showed that the numeric target is by far the largest reason for that gap. Removing it alone accounts for most of the difference, while the other additions contribute much less. That result is real, but it does not test the part of the system this paper is about. The direction-reading step is not exercised at all when a scene is generated with no user direction, yet it is present in the full comparison. So the full-system number should be read as evidence that the numeric targets are followed, not as evidence that direction-reading works. To test direction-reading on its own, we hold everything else fixed and turn only that step on and off.

Isolating the direction-reading step

We draw seven scenes from a single feature-length test screenplay of 118 scenes, *Hamnet*, which dramatizes the death of Shakespeare’s eleven-year-old son and its effect on his mother Agnes and the rest of the family. The seven scenes span roughly the first tenth of the story to the last eighth, so the result does not depend on one part of it. We fix a panel of six plain-language directions, held constant across all seven scenes: make it more tense and let the confrontation escalate; make it quieter and slow the pacing; raise the stakes; keep it calm and mostly dialogue; give it more physical action and less talking; and make the tone lighter. For each scene we generate under two conditions, with everything but the direction held fixed, including the story context and the sampling temperature. Condition A skips the direction-reading step entirely and generates against the scene’s plain target; Condition B reads the direction into changes, clips and applies them, and then generates. The writing step itself is performed by Claude Haiku 4.5: the model is prompted with the story context and the six target numbers, drafts a scene, and then revises its own draft. Only that step changed. Direction-reading is the temperature-zero Llama-3.3-70B-Instruct call of the control loop, and target derivation and scoring reuse the deterministic

Measure	Sign agreement	Pearson r
Tension	65.9% (27/41)	+0.30
Mood	85.7% (6/7)	—
Volatility	75.0% (30/40)	+0.36
Pacing	58.5% (24/41)	+0.27
Dialogue/action ratio	61.5% (8/13)	+0.05
Peak intensity	53.3% (8/15)	-0.17

Table 3: Directional agreement per measure ($K=4$ reps). Mood is a separate follow-up.

analyzer; none of these depends on which model does the writing. One (scene, direction) unit is excluded throughout this section: three of its four repeated generations came back empty after a long provider-side stall, leaving only one usable draw rather than the four every other unit rests on, so we drop it rather than average over data that was never really there.

A single draw from either condition is noisy: the same request can move a measure by a small amount in either direction depending on how the model happens to phrase a line, which can hide a real effect or manufacture one that is not there. To let a genuine effect surface above that noise without inventing one, we generate each condition four times independently and average the resulting measured vectors before comparing them, rather than judging a direction from a single pair of scenes. This averaging method was fixed in advance of scoring, not chosen after seeing which scenes looked best.

We score every generated scene with the deterministic analyzer and average the four measured vectors per condition to get v_A and v_B . For each measure that the direction actually changed, that is, each measure with a non-zero requested change, we compare the direction of the requested change with the direction of the measured change $v_B - v_A$. The headline number is the *directional agreement rate*: the share of (scene, direction, measure) cases where the two agree in sign. As a secondary check that also reflects magnitude, we report per measure the Pearson correlation between the requested change and the measured change across its (scene, direction) cases. We also compare $D(v_A, t_A)$ with $D(v_B, t_B)$ to check whether reading a direction pushes the scene away from its target, that is, whether the added step costs compliance.

Table 3 gives the result: five of six measures move in the requested direction clearly above the 50% chance level, volatility most strongly of all, on the exact measure that showed no effect whatsoever in every earlier, smaller round of this same ablation (the harder-measures analysis). Peak intensity is the exception, and adding a seventh scene made it weaker, not stronger, dropping from 61.5% on six scenes to 53.3% on seven, barely above chance. Part of this is a property of the panel rather than the measure: across all seven scenes, the six directions in this panel only ever move the peak-intensity target by two distinct amounts, +0.2 or -0.1, so a correlation computed against that request has almost no spread to work with and is easily moved by a single case; the sign-agreement figure is the more trustworthy of the two numbers here, and even it does not clear chance convincingly. It is a genuinely weak result, and the harder-measures analysis traces part of it to the same kind of directive-wording gap already closed for tension. Compliance moved the other way from directional agreement. Averaged across all 41 (scene, direction) units, $D(v_A, t_A) = 1.37$ and $D(v_B, t_B) = 2.25$: guided generation lands, on average, 1.6× farther from its own target than the

undirected baseline lands from its own. This is not a new finding on its own; every earlier, smaller version of this ablation showed the same direction of effect, between 1.1× and 1.8×. What is new is that the gap has now held up under a larger, more varied sample, and we return to a possible reason for its size in the limitations.

Collapsing Table 3 the other way, across measures instead of across directions, shows the same six directions are not equally reliable. “Make it more tense and let the confrontation escalate” agrees with the requested direction in 74% of its (scene, measure) cases and “give this more physical action, less talking” in 71%, the two most reliable directions in the panel. “Make the tone lighter here” is the least reliable, at 43%, well below every other direction and the only one below the 50% chance level; the harder-measures analysis traces why this specific direction was the hardest to satisfy.

To make the numbers concrete, the two passages below show the kind of difference Table 3 measures. Both are generated by our system for one scene of the *Hamnet* screenplay [9]; neither is taken from an existing film or script. The baseline and the “more tense” condition share the scene, characters, and story position, and differ only in the direction given.

Condition A (no direction):

Hamnet, eleven, stands alone with a wooden sword, jabbing at shadows. His movements are jerky, uncertain. He stops. Listens to nothing.

AGNES: What are you killing?

HAMNET: Nothing. Everything.

Condition B (“make it more tense and let the confrontation escalate”):

She finds Hamnet at the far corner, crouched beside a crude wooden structure, boards lashed together with twine. He doesn’t notice her. He’s holding a wooden sword, moving through the space with deliberate steps, his mouth working silently.

AGNES: What are you building?

Hamnet startles. The sword drops.

Both scenes open on the same boy with the same wooden sword; the guided version replaces an interior, unwitnessed moment with one that is interrupted and reactive, the change the direction asked for without the target or the setting being restated anywhere in the prompt beyond the six numbers and the arc label behind them.

Diagnosing and fixing the harder measures

Not every measure was equally easy to steer, and the fixes ranged from rewording a directive to correcting how a target was computed in the first place. Pacing’s difficulty had a concrete, fixable cause: the valid range described in the control loop once allowed target values the underlying formula can never produce, so some requests asked for a pacing figure that was, by construction, unreachable. Correcting that range, together with giving tension the same kind of translated writing instruction the other measures already had, recovered directional control for both, and those fixes are reflected in Table 3.

Volatility needed a fix one level further back, in how its target was derived from the corpus in the first place, not in how the direction was worded. Measured across the full training corpus, a scene’s raw sentiment-swing statistic rises with how many lines it has, from about 0.12 at two or three lines to about 0.22 at thirty or more, independent of whether the scene itself reads as calm or turbulent, a property of how

a standard deviation behaves on a small sample, not of the writing. Three of the six scene-type centroids that supply the target-derivation stage's baseline targets carried a volatility of exactly zero, the value the underlying formula falls back to when a scene has fewer than two analyzed lines, not a genuine signal that those scene types are calm. Every scene our system generates carries enough dialogue to sit well above that length, so asking a new scene to match one of these targets set a goal close to unreachable, whatever the direction said. We re-derived the six scene-type centroids on a length-normalized version of volatility, the raw statistic divided by its corpus-wide expected value for a scene of that length, then converted the resulting centroids back onto the original scale using the typical length of a generated scene, so nothing downstream of the target itself needed to change. Table 3 reflects this correction, and volatility moves from the weakest measure in earlier, smaller rounds of this ablation to the strongest.

Peak intensity's own definition, the scene's single most intense physical beat rather than a scene-wide average, made it more sensitive than the other five measures to a mismatch between what its directive described and what its scorer actually rewards. Checking the analyzer directly against short hand-written test lines showed that this score comes from a small lexicon of concrete verbs, slams, grips tightly, smashes, collapses, screams, rather than from how much physical activity a passage describes: an unambiguous but abstract line like "the fight breaks out violently, chaos erupting around them" scored only 0.55, barely above a scene of someone sitting quietly (0.475), while a single line built around "slams" or "grips tightly" scored 0.99 regardless of the rest of the scene. The directive in place during the main study asked for a "violent or urgent burst of physical action" without naming any of the vocabulary the score actually responds to, the same kind of gap already closed for tension (translated into concrete confrontation acts) and volatility (told where its signal lives, in dialogue rather than action lines). Naming the same small set of decisive verbs directly in the directive made intensity requests move the right way far more often, though, as the uniform panel shows in detail, the fix reliably raises intensity on request while lowering it stays harder for a reason built into the measure. Peak intensity's figure in Table 3 reflects the directive used during the main study, before this fix.

Mood needed two further corrections, both found by reading generated scenes rather than by inspecting a formula. The band of values used to turn a numeric mood target into a written instruction, dark, neutral, or bright, had been calibrated to a wider range than this corpus actually produces, so almost every scene, whatever the direction, was given the same "neutral" instruction; narrowing that band let the instruction vary with the target at all. Once it did, a request for a brighter scene was still sometimes met with a different kind of dark material, an illness, a death, a folk ritual, rather than open conflict; naming that pattern directly, alongside concrete, ordinary alternatives such as food, comfort, or small successes, removed most of it. A related case remained: scenes whose established situation was already serious, tending a wound, for instance, were still written with a matching clinical seriousness even once the darker substitutions were gone. Asking instead for warmth in how the characters handle a serious situation, not around it, resolved the rest. Because these two fixes were made after the run behind the first six scenes of Table 3, mood's figure there combines two sources: the six original mood-touching (scene,

Measure	Sign agreement	Raise	Lower
Volatility	75% (6/8)	2/4	4/4
Pacing	71% (5/7)	1/3	4/4
Dialogue/action ratio	67% (4/6)	2/4	2/2
Mood	62% (5/8)	3/4	2/4
Tension	50% (4/8)	2/4	2/4
Peak intensity	50% (4/8)	3/4	1/4
Overall	62% (28/45)		

Table 4: Uniform panel: a raise and a lower direction per measure, four scenes, $K=4$ reps.

direction) cases, regenerated under the corrected directive and compared against the same baselines already measured, rather than a full new run across all six directions; and the seventh scene's own "make it lighter" case, generated once under the same corrected directive alongside its other five directions. Mood's directional agreement on the six-scene check rose from 33% before these two fixes to 50% after the first and 83% after the second; the seventh scene's case agreed as well, giving 6/7 overall.

A uniform per-measure panel

The six directions in the main panel are the kind of thing a screenwriter actually types, which is their strength, but it also means they do not exercise the six measures equally. A plain direction changes only the measures it implies, so "make the tone lighter" never asks anything of the dialogue-to-action ratio, and the sample behind each measure in Table 3 runs from forty-one cases down to seven. To compare the six on equal footing, we ran a second, deliberately controlled panel: two directions per measure, one asking to raise it and one to lower it, each written around that measure alone, over four scenes spanning the story, with the same $K=4$ reps and the same generation and scoring machinery as before. Because these directions are narrower and more artificial than the ones a writer would actually type, this panel is a controlled diagnostic rather than a test of realistic use.

Table 4 gives the result; a few cells are dropped where the parser left the target measure unchanged, so some measures have fewer than eight cases. Across all six measures the panel agrees with the requested direction in twenty-eight of forty-five scored cases (62%). Volatility again steers among the most reliable of the six, and the compliance cost from the main study reappears here at 1.3 $\times$. Two measures come in at chance, for different reasons.

Peak intensity's overall figure hides a clean split: raising it succeeds in three of four scenes, while lowering it succeeds in only one. This follows from how the measure is defined. It is the single most intense action beat in a scene, a maximum rather than an average, so one decisive verb is enough to raise it, but lowering it means holding every beat below that level at once. The directive fix from the harder-measures analysis, which names the vocabulary the score responds to, is what makes the raising direction reliable; the lowering direction stays hard for a reason that sits in the measure, not the prompt. Tension, reliable in the main study at 66%, drops to chance here. Its two directions ask directly for sharper or softer confrontation, a narrower and more forceful demand than the main panel's "make it more tense," and across eight cases the difference between reliable and chance is a single scene. We read this as a caution about how far one measure's number moves with the exact wording of a request, not as evidence

Dimension	Agreement	p
Dialogue vs. action	100% (28/28)	< 0.001
Physical action	90% (38/42)	< 0.001
Tension	61% (34/56)	0.14
Overall	79% (100/126)	$< 10^{-4}$

Table 5: Reader agreement with the intended version, nine curated pairs, fourteen readers, two-sided exact binomial against chance.

that tension is uncontrollable, given the larger main-panel sample that says otherwise.

HUMAN PERCEPTION OF THE CHANGES

The control loop shows that our analyzer registers the change a direction asks for. It does not show that a person reading the scene would notice it. The two can come apart: a measure can shift for a reason a reader never registers, such as a single word the score reacts to, so a scene that satisfies the analyzer need not read any differently to a person. Whether readers actually feel these changes is a separate and harder question, and we ran a small study to answer it.

For material, we generated about fifty pairs of scenes, each a plain baseline against a directed version, across all six measures in both directions and about ten screenplays outside the corpus, and scored every pair with the same analyzer.¹ Most pairs did not end up with a difference a reader could point to, and the reasons lie in the measures, not the writing. Peak intensity is the single strongest beat in a scene, so a request for more action is often met by one sharp moment late in the scene while the opening the reader actually reads is unchanged; on one pair the whole-scene value rose from 0.65 to 0.99 while the first two hundred words held flat at 0.55. The same measure can rise on language that only sounds intense, scoring a character alone at a desk on the phone near the top of its range. And the baseline is rarely calm to begin with: the undirected model already writes tension into a dramatic scene, so the directed version is not always a clear step above it.

So we ran the study on the pairs where the question is fair: the ones where the intended change is clearly there in the passage the reader sees. We chose those pairs by reading the text, before collecting any answers, and we list them in full below so the choice is open to inspection. Reader answers were used only to score the pairs, never to pick them. The final set is nine pairs across three dimensions anyone can judge without training, physical action, tension, and how much a scene is dialogue rather than action, taken from six screenplays. Fourteen readers saw each pair in a random left-right order and chose the version that better fit a short question, such as which one has more physical action.

Table 5 shows the result. Overall, readers picked the intended version in 79% of the 126 judgments, well above the 50% of a coin ($p < 10^{-4}$). The breakdown by dimension is the more useful number. The two concrete dimensions, physical action and the dialogue-to-action balance, were read almost perfectly, at 90% and 100%. Tension was mixed: 61% overall, above chance but not enough to be sure at this sample size.

¹The models used earlier in this work were no longer available for these later runs; direction-reading and generation here both use Claude Haiku 4.5. The measures used to pick pairs, physical-action intensity and dialogue and action line counts, are computed locally without a language model.

The reason is clear pair by pair. When the tension is concrete, a pistol at a captive's head or a hand on a sword in a standoff, twelve of fourteen readers picked it, as reliably as the action pairs. When it is only a sharper line of dialogue against an already-tense scene, readers were split. We report all nine pairs as chosen, the two weak ones included, so the numbers reflect the set we fixed in advance and not one trimmed to fit. This matches the structural results of the validation on the reader's side. The changes readers notice are the concrete, countable ones, which are also the ones the analyzer steers best. Abstract emotional change is harder in both places, to steer and to notice. The clear finding is the positive one: a concrete structural change is not just measured, it is felt. The abstract case stays open. Tension still leans above chance overall, its concrete forms come through as well as physical action, and only its dialogue-only form is undecided, resting on two pairs, too few to say whether a soft change in tension can be felt at all. A larger, more sensitive study is the way to settle that, and this work makes the question precise enough to ask.

LIMITATIONS

This evaluation checks the direction of change against a fixed panel of six directions on our own test scenes. It does not check that the change the system reads matches what a screenwriter giving that direction actually meant. A phrase like "raise the stakes" has no single correct numeric reading. The panel and its ranges are one reasonable interpretation, and comparing the system's reading against practicing screenwriters' own intent is left for future work. The distance in Equation 2 is a measure of compliance, not of quality, and our earlier finding that compliance tracks judged quality rested on a single test project, too small to support a general claim. The corpus leans toward drama and comedy, and reducing a scene to six numbers sets aside subtext, metaphor, and physical comedy. A direction whose effect lives in one of those dimensions would not show up in our measurements at all, whether or not it worked.

The reader study speaks to whether a change is felt and not only measured, but it does so on pairs chosen because the change is plainly present in the text. It shows that concrete changes are perceived; it does not measure how often an arbitrary structural change would be. Mapping where the line between a perceptible and an imperceptible change falls, and how far the tension result holds up under a larger sample, is the natural next study.

For mood and volatility, the scoring used to check a generated scene against its target is still not computed identically to the statistic used to build the corpus-derived archetypes, even after the harder-measures analysis's fixes. The archetype-level mood figure gives most of its weight to how a scene ends, while our scoring of newly generated text averages sentiment across the whole scene; the archetype-level volatility figure, even length-normalized, is computed over every line, while our scoring restricts the swing calculation to dialogue lines, where the signal actually lives. The corpus-derived targets themselves are untouched beyond the length correction described in the harder-measures analysis; only the way the ablation scores newly generated text differs from how the original archetypes were built. Mood's reported figure is also the thinnest in Table 3: seven cases, all from a single direction, six of them regenerated separately from the other five

measures once its directive was corrected, rather than the full, larger sample the other measures rest on.

The compliance-cost gap (D_B landing farther from target than D_A) is the one finding present in every configuration of this ablation we have run. Its size, though, is not stable: it was largest on six scenes (1.8 $\times$) and fell back to 1.6 $\times$ once we added a seventh. This suggests that part of the gap's size, though not its direction, may be an artifact of how tightly the undirected baseline happens to fit its own target, rather than a true cost of reading a direction. Across all seven test scenes, the scenes whose baseline already sat closest to its target showed the largest gap when a direction was added, while scenes whose baseline fit more loosely showed a much smaller gap, in one case close to none at all; this pattern held, and strengthened, with the seventh scene added ($r = -0.70$ between a scene's baseline distance and its gap, up from $r = -0.62$ on six scenes). A baseline that already sits almost on its target can only move farther from it once the scene changes at all, even when the change is the one the direction asked for. This does not remove the finding, which has held in every run regardless of scene, model, or the fixes described in the harder-measures analysis, but it means the honest interpretation is narrower than "asking for a direction makes the scene worse": it may say more about how compliance is measured at a single point than about the direction-reading step itself. We flag this as a concrete question for future work rather than resolve it here.

CONCLUSION

Current language-model co-writing tools let a screenwriter state a direction in plain language but give no way to see whether the tool acted on it. We described a system that reads such a direction into small, bounded changes over a structural target learned from a corpus of screenplays, and that reuses the same analyzer which produced the target to measure the finished scene. This closes the loop between asking for a change and checking that it happened. We also isolated the direction-reading step with a direct ablation, rather than leaning on a full-system comparison that a larger part of the pipeline dominates. Doing so surfaced two kinds of problem beyond wording alone: a target computed from the corpus in a way that made part of it nearly unreachable, and a direction reliably read but not reliably followed in writing. Fixing both brought five of the six structural measures we track to a directional agreement rate clearly above chance, between 58% and 86%. The sixth, peak intensity, was the weakest, and a controlled follow-up showed why: a request to raise physical intensity is followed reliably once the directive names the cues the measure responds to, while a request to lower it runs up against how the measure itself is defined. A reader study on the finished scenes then asked whether these changes reach a person and not just the analyzer: readers identified the concrete changes, physical action and the balance of dialogue against action, almost perfectly, while abstract tension nuance stayed at chance, mirroring on the reader's side the

same split between concrete and abstract measures that the analyzer shows. The cost of reading a direction at all, a scene that lands farther from its own target than an undirected one lands from its own, persisted through every version of this test regardless of which problem we fixed or how many scenes we tested it on, and remains the clearest open question this work leaves behind.

References

- [1] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.
- [2] Syd Field. 2005. *Screenplay: The Foundations of Screenwriting*. Delta.
- [3] Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 14504–14528.
- [4] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. arXiv:1909.05858.
- [5] Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable Text Generation for Large Language Models: A Survey. arXiv:2408.12599.
- [6] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.
- [7] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–34.
- [8] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2241–2252.
- [9] O'Farrell, Maggie and Chloé Zhao. 2025. *Hamnet*. Unpublished screenplay.
- [10] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie Plot Analysis via Turning Point Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 1707–1717.
- [11] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. 2020. Screenplay Summarization Using Latent Narrative Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1920–1933.
- [12] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science* 5, 1 (2016), 31.
- [13] Blake Snyder. 2005. *Save the Cat! The Last Book on Screenwriting You'll Ever Need*. Michael Wiese Productions.
- [14] Yunge Wen, Chenliang Huang, Hangyu Zhou, Zhuo Zeng, Chun Ming Louis Po, Julian Togelius, Timothy Merino, and Sam Earle. 2025. All Stories Are One Story: Emotional Arc Guided Procedural Game Level Generation. arXiv:2508.02132.
- [15] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating Longer Stories With Recursive Reprompting and Revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4393–4479.
- [16] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-Write: Towards Better Automatic Storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7378–7385.
- [17] Rongzhi Zhang, Liqin Ye, Yuzhao Heng, Xiang Chen, Tong Yu, Linghai Kong, Sudheer Chava, and Chao Zhang. 2025. Precise Attribute Intensity Control in Large Language Models via Targeted Representation Editing. arXiv:2510.12121.
- [18] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
- [19] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*.