Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Prompt Learning with Vision--Language Foundation Models for Automated Storyboard Assessment and Formative Feedback in Film Education

Lu Zhang

zhangluhefive@163.com

Chongqing Technology and Business University

Research Article

Keywords: storyboard assessment, vision–language foundation models, multimodal prompt learning, ordinal rubric modeling, counterfactual feedback, optimal transport, film education

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11412390/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Prompt Learning with Vision–Language Foundation Models for Automated Storyboard Assessment and Formative Feedback in Film Education

Anonymous Author[1,2*], Anonymous Author[2,3†] and Anonymous Author[1,2†]

[1*] Department of [Anonymous], [Anonymous University], [Anonymous], [Anonymous], 000000, [Anonymous], [Anonymous] .
[2] Department of [Anonymous], [Anonymous University], [Anonymous], [Anonymous], 000000, [Anonymous], [Anonymous] .
[3] Department of [Anonymous], [Anonymous University], [Anonymous], [Anonymous], 000000, [Anonymous], [Anonymous] .

*Corresponding author(s). E-mail(s): anonymous@review.com;
Contributing authors: anonymous@review.com; anonymous@review.com;
†These authors contributed equally to this work.

Abstract

Storyboards articulate cinematic meaning through compositions, actions, captions, and the temporal relations between shots. Assessing them is challenging because recognizable shot attributes do not uniquely determine narrative quality, and persuasive feedback may lack evidence in the submitted panels. We formulate storyboard assessment as criterion-conditioned inference over an editable temporal graph and propose Criterion-constrained Rubric-aware Feedback Transport (CRaFT). The approach couples frozen vision–language encoders with low-rank criterion prompts, unbalanced optimal transport between multimodal evidence and rubric anchors, and ordinal score prediction. A graph-consistent counterfactual objective scores candidate revisions by their localized, criterion-specific effects, while conservative selection can abstain when support or uncertainty is inadequate. On the reported public-data evaluations (MovieShots, FullShots, AVE, and MovieNet), CRaFT obtains macro-F1 scores of 77.4%, 73.1%, 72.3%, and 70.0%, respectively, exceeding the strongest listed prompt baseline by 5.2–5.9 percentage points across these tasks. The reported AVE sequence-order accuracy reaches

80.1%, while the mean ordinal agreement on constructed rubric proxies is 0.69 quadratic-weighted kappa. All evaluated source datasets are previously released public cinematic resources. Constructed rubric-proxy scores and controlled editing perturbations assess observable cinematic structure and model behavior; no students or instructors were recruited and no teacher-rated storyboard validation was performed. The experiments therefore cannot establish instructor agreement, teacher endorsement, or improved student learning. This distinction between computational proxy evaluation and future educational validation bounds the conclusions.

Keywords: storyboard assessment, vision–language foundation models, multimodal prompt learning, ordinal rubric modeling, counterfactual feedback, optimal transport, film education

1 Introduction

Storyboards are compact representations of cinematic intent: a panel encodes framing, camera position, subject arrangement, staging, action, and anticipated transitions to neighboring panels. In film education, instructors assess more than the correctness of these visual attributes. They must judge how camera grammar supports an intended story, whether screen direction remains legible, and which revisions could help a student convey an emotion or event more clearly. This process is contextual and iterative. A single close-up may be technically well composed but narratively unmotivated; an apparent violation of continuity may equally be a deliberate stylistic decision. Large public resources such as MovieNet, MovieGraphs, MovieShots, and the Anatomy of Video Editing (AVE) dataset provide valuable supervision about cinematic structure and shot attributes, but they are not designed as annotated collections of students' formative feedback [1–4]. The central opportunity is therefore not to replace film instructors with an automatic taste score, but to produce auditable, criterion-specific assistance that makes the basis of a judgment explicit.

Recent vision–language systems offer a rich starting point. Contrastive models align visual evidence and language descriptions [5–7], while multimodal instruction-tuned systems can describe detailed visual scenes [8–11]. Parameter-efficient adaptation through learned context tokens has progressed from CoOp, CoCoOp, and visual prompt tuning to coupled multimodal prompts and self-regularized prompt optimization [12–17]. In parallel, video understanding models have improved the representation of long narratives, temporal interactions, and shot-level semantics [18–20]. Yet the direct use of these tools for teaching frequently takes a linear form: encode an image, assign a score, and prompt a language model to explain it. Such a pipeline can yield plausible feedback without identifying the piece of evidence on which the criticism rests or establishing that a suggested edit addresses the cited criterion.

Three obstacles prevent this straightforward adaptation from supporting reliable assessment. *First, criterion entanglement*: framing, continuity, narrative intelligibility, and pacing often share visual cues, yet a fixed global embedding may not preserve

the distinctions required by a multi-criterion rubric. *Second, counterfactual ambiguity*: descriptive recognition of a reversed eyeline, for example, does not establish whether changing it would improve the storytelling goal; many filmic conventions admit purposeful exceptions. *Third, validity and uncertainty*: public movie data encode professionally edited outcomes, not instructor judgments of learner work, and may bias a model toward imitating the dominant cinematic style. Spurious certainty is especially problematic when the learner's creative intention, motion, or audio is unavailable. Recent findings about video-model hallucination, temporal inconsistency, and criterion-sensitive automated grading further motivate evidence checks, selective feedback, and independent educational validation [21–24].

We address these issues by treating a storyboard as an *editable, criterion-labeled sequence graph* and learning prompts through the consequences of graph-feasible perturbations. The proposed Criterion-constrained Rubric-aware Feedback Transport (CRaFT) learns a transport plan between scene evidence and explicit rubric anchors, a monotone ordinal score on each criterion, and an edit-sensitive value function in one constrained objective. Instead of appending an explanation generator to a classifier, a candidate feedback action must be associated with a localized evidence path, a supported intervention, and a conservative lower bound on estimated improvement. The foundation encoders remain frozen, allowing controlled adaptation without training a complete multimodal model. This design is intended to yield interpretable decisions that can be inspected and overruled by teachers, not universally valid aesthetic grades. It also motivates a two-level evaluation: public cinematic datasets for observable attributes and structured perturbations on existing publicly released film datasets. No instructor-labeled or student-participant evaluation was conducted; independent educator validation and longitudinal learning outcomes remain future work. Our contributions are summarized as follows.

1. We formulate criterion-based storyboard assessment as *rubric-constrained graph inference*, distinguishing observable cinematic attributes, pedagogical judgments, and declared creative intent.
2. We introduce CRaFT, which couples multimodal prompt adaptation, unbalanced evidence-to-rubric transport, and monotone ordinal assessment within a single differentiable objective rather than independent recognition and grading modules.
3. We derive an uncertainty-constrained counterfactual feedback rule in which every suggested edit has an explicit graph operator, evidence attribution, and testable lower-confidence improvement condition; we state precisely which guarantees rely on predictive error assumptions.
4. We specify reproducible shot-recognition, intervention, calibration, and evidence-localization protocols based on published film annotations and explicit computational proxies. External instructor assessment and student-learning effectiveness remain outside the present experimental scope.

2 Related Work

2.1 Film understanding and cinematic attribute modeling

Earlier movie understanding benchmarks concentrate on situations and story structure rather than educational judgment. MovieGraphs annotates human-centric relations and interactions in movie clips [2]; MovieNet covers scenes, shots, style-related tags, and multiple modalities [1]. LSMDC-style movie descriptions facilitate narration and retrieval [25]. MovieShots and MovieShots2 study the semantics of shot scale, movement, and composition [3]; AVE extends coverage to shot attributes, camera setups, and editing transitions [4]. Subsequent cinematic classification research introduced FullShots and compared motion-sensitive and multi-task architectures across shot properties [26, 27]. These resources suggest a useful separation between factual labels such as "medium shot" and rubric-level claims such as "the sequence establishes location clearly." We use the former for observable-proxy supervision while treating the latter as a distinct evaluation target requiring dedicated annotation. Other movie and long-video models—MovieChat, Video-LLaVA, Video-MME, LongVideoBench, and MLVU—highlight the challenge of temporally dispersed evidence [18, 28–31].

2.2 Multimodal foundation models and prompt adaptation

CLIP introduced language-supervised transferable visual representations, while SigLIP and EVA-CLIP explored alternative objectives and scaling [5–7]. BLIP-2, InstructBLIP, and visual instruction tuning provide richer generative vision–language interfaces [9–11]. ImageBind, Segment Anything, DINOv2, Grounding DINO, and Florence-2 extend the range of region, modality, and grounding evidence available to downstream models [32–36]. Prompt learning has moved from text-only or conditional context optimization to joint visual/text prompts, self-regularization, knowledge preservation, missing-modality handling, and interactive fusion [12, 13, 15–17, 37, 38]. More recent work studies prompt distillation, attribute-conditioned prompts, meta-regularization, black-box search, multitask transfer, and prompt-free alternatives [39–46]. These papers chiefly evaluate representation transfer, not the logical compatibility of generated feedback with a film rubric. Our novelty claim is therefore a *task formulation and a coupled optimization*, not the invention of prompt tokens in isolation.

2.3 Temporal grounding, educational feedback, and model reliability

Video temporal reasoning increasingly relies on memory, hierarchical search, and query-conditioned selection [19, 20, 47–50]. Hallucination and instability remain substantial concerns in video question answering and grounding [21, 22, 51, 52]. Automated educational grading research indicates that even fluent feedback must be evaluated for criterion agreement, stability, and learner agency [23, 24, 53]. We retain the useful representation-learning principles of these systems while requiring a direct path from evidence to a criterion and then to a proposed edit. We do not equate agreement with automatically derived labels to validity against instructor judgments.

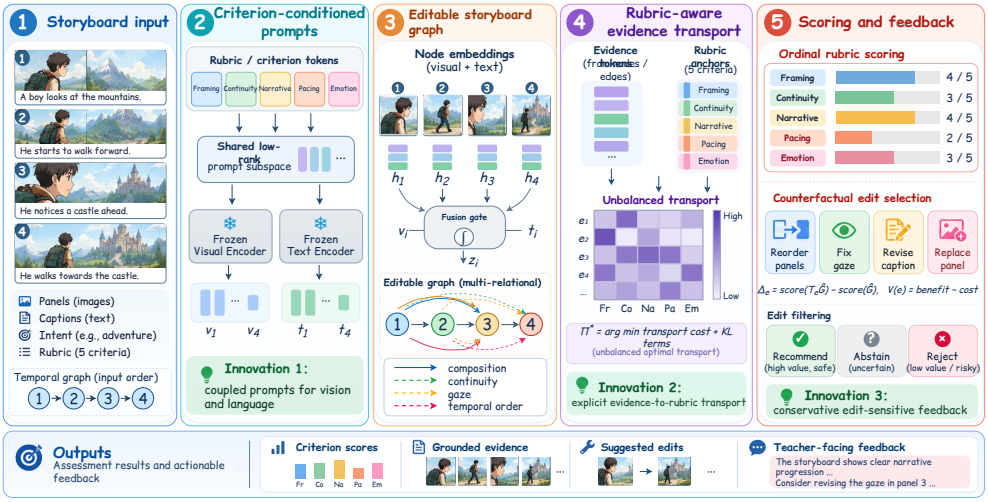


Fig. 1 Overall architecture of the proposed CRaFT framework. Given a storyboard sequence with visual panels, captions, creative intent, and assessment rubrics, CRaFT performs criterion-conditioned multimodal representation learning through a shared low-rank prompt subspace and frozen vision–language encoders. The resulting representations are organized into an editable temporal graph that captures cinematic and narrative relationships. Unbalanced optimal transport aligns multimodal evidence with criterion-specific rubric anchors, enabling grounded ordinal assessment. Finally, counterfactual edit evaluation estimates criterion-specific improvements and selects feasible revisions under evidence and uncertainty constraints. The framework jointly produces rubric scores, localized evidence, and actionable formative feedback.

Human interpretation and feedback validity. Representation discrepancies can be treated as prompts for human interpretation rather than automatic judgments; this distinction is illustrated by Rafaeli and colleagues' multi-layer literary-network analysis [54]. In film education, the formative role of evidence-linked comments is informed by Hattie and Timperley's model of feedback [55]. Neither analogy validates the present rubric or demonstrates learning gains, which require separate instructor and learner studies.

3 Method

3.1 Task definition, editable graph, and observability

Let an input storyboard be $\mathcal{B} = \{(I_i, c_i, z_i, t_i)\}_{i=1}^{N}$, where I_i is a sketch or selected keyframe, c_i its optional caption or dialogue, z_i optionally describes camera scale, angle, and motion, and t_i supplies a relative scene position or shot duration. An instructor may additionally provide intent g (for example, "unsettling discovery"), constraints q, and a rubric $\mathcal{R} = \{r_k\}_{k=1}^{K}$, each with an ordered set of $L_k{+}1$ levels. Missing modalities are explicitly represented by a mask m_i; null captions are never replaced by generated text and silently treated as observed.

We construct an editable attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$, with panel nodes, temporally directed edges $(i, i+1)$, and optional higher-order motif edges (i, j). Each node

contains a fused representation and each edge stores geometrical, identity, continuity, and narrative relation estimates. A *legal edit* $e \in \mathfrak{E}(\mathcal{G}, g, q)$ is an operator $T_e : \mathcal{G} \mapsto \mathcal{G}_e$ that changes an allowed attribute, revises a caption, replaces a panel, or modifies an order relation subject to the user's declared constraints. We separate $\mathcal{R} = \mathcal{R}_{\text{obs}} \cup \mathcal{R}_{\text{ped}}$: observable criteria admit external categorical labels, while pedagogical criteria require instructor scoring before reliability claims. This distinction is used in both learning and evaluation.

Canonical criteria and scoring scale. In the computational rubric-proxy evaluation, the five rubric-based criteria are *composition, continuity, narrative clarity, legibility, and pacing*, each on the ordinal scale $\{0, 1, 2, 3, 4\}$. Composition incorporates framing and subject arrangement; legibility concerns readability of the intended action and screen direction; narrative clarity concerns whether panel transitions convey a coherent causal progression given the stated intent. Pacing concerns relative temporal emphasis only when sequence or duration evidence is observable. Emotional intention is contextual input g, not a sixth independent scoring dimension. The terms "framing" and "emotion" appearing in Figs. 1 and 2 describe model inputs, not the final rubric taxonomy. The 1–5 marks in the schematics indicate display indices $d_k = \hat{s}_k + 1$; all quantitative proxy analyses use the 0–4 scale.

Rubric anchor operationalization. Level 0 denotes missing or strongly contradictory observable support; level 1 denotes major ambiguity; level 2 denotes mixed support; level 3 denotes largely interpretable criterion evidence; level 4 denotes clear, sustained support for the stated creative intent. These generic descriptors must be supplemented by the actual criterion-by-level anchor phrases supplied to annotators (Rubric-v0.1, comprising five 0–4 criteria and 25 operational anchor descriptors in Table 2; these anchors define computational proxies rather than instructor-assigned classroom grades). A technically unconventional shot is not automatically penalized when its intended function is documented.

Given $\mathcal{G}$, the desired system returns criterion scores $\hat{s}_k \in [0, L_k]$, confidence sets $\mathcal{C}_k$, grounded evidence $\mathcal{A}_k \subseteq \mathcal{V} \cup \mathcal{E}$, and optionally an edit set $S \subseteq \mathfrak{E}$ with a natural-language explanation of each selected action. The objective is not to maximize the grades of standard feature films: it is to estimate a rubric-consistent assessment of the submitted representation and propose verifiable candidates under uncertainty.

The overall architecture of CRaFT is illustrated in Fig. 1, which integrates criterion-conditioned multimodal prompting, editable temporal graph modeling, rubric-aware evidence transport, and counterfactual feedback generation.

3.2 Criterion-conditioned multimodal prompt geometry

Let E_v, E_t be fixed visual and text encoders. Criterion keys $u_k \in \mathbb{R}^{d_u}$ generate coupled visual and linguistic prompt tokens through a common low-rank latent factor $H_\ell \in \mathbb{R}^{r \times d_u}$. For m tokens per branch and encoder widths d_t, d_v, let $C_\ell^t \in \mathbb{R}^{(md_t) \times r}$ and $C_\ell^v \in \mathbb{R}^{(md_v) \times r}$. At layer ℓ, define

$$P_{k,\ell}^t = P_{0,\ell}^t + \text{mat}_{m,d_t}(C_\ell^t H_\ell u_k), \tag{1}$$

$$P_{k,\ell}^v = P_{0,\ell}^v + \text{mat}_{m,d_v}(C_\ell^v H_\ell u_k), \tag{2}$$

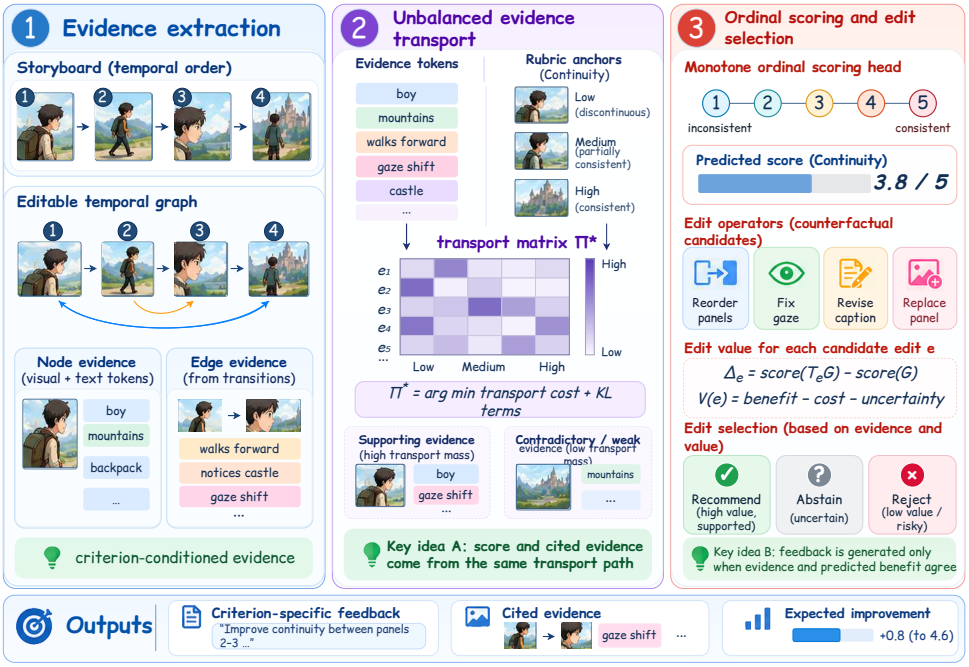


Fig. 2 Detailed illustration of rubric-aware evidence transport and counterfactual feedback in CRaFT. First, criterion-conditioned node and edge evidence is extracted from the editable storyboard graph, capturing visual attributes, textual semantics, and temporal relationships. Next, unbalanced optimal transport establishes soft assignments between evidence tokens and ordered rubric anchors while allowing unsupported evidence to remain unmatched. The resulting transport plan provides both criterion-specific evidence attribution and representations for monotone ordinal scoring. Finally, candidate storyboard edits are evaluated by recomputing rubric scores on modified graphs. Evidence support, predicted improvement, and uncertainty jointly determine whether an edit is recommended, rejected, or withheld. The illustrated scores and revisions are schematic examples.

$$\operatorname{rank}(H_\ell) \le r \ll d_u. \tag{3}$$

Here $\mathrm{mat}_{m,d}$ reshapes a vector into m prompt tokens of width d. Sharing H_ℓ means that criterion-specific changes in visual and linguistic prompts originate in the same low-dimensional rubric subspace, rather than in independently fitted modules. The common key u_k conditions a prompt on a real rubric description, examples of acceptable evidence, and observed metadata; it is not a latent label for an immutable aesthetic style. For each panel i and criterion k, produce

$$\boldsymbol{v}_{ik} = \mathrm{norm}\left[E_v(I_i; P_k^v)\right], \tag{4}$$

$$\boldsymbol{h}_{ik} = \mathrm{norm}\left[E_t(c_i \oplus r_k \oplus g; P_k^t)\right]. \tag{5}$$

An availability-aware gate avoids inventing missing evidence:

$$\alpha_{ik} = \frac{m_i^v e^{a_v(\boldsymbol{v}_{ik})}}{m_i^v e^{a_v(\boldsymbol{v}_{ik})} + m_i^t e^{a_t(\boldsymbol{h}_{ik})} + \epsilon}, \tag{6}$$

$$\boldsymbol{x}_{ik} = \alpha_{ik} m_i^v \boldsymbol{v}_{ik} + (1 - \alpha_{ik}) m_i^t \boldsymbol{h}_{ik}. \tag{7}$$

where the absent-branch vector is masked before fusion. To represent relations between panels, we use a structured message passing rule indexed by cinematic relation type ρ:

$$\boldsymbol{h}_{i,k}^{(s+1)} = \mathrm{LN}\left[\boldsymbol{h}_{i,k}^{(s)} + \sum_{j \in \mathcal{N}(i)} \eta_{ij,k}^{(s)} W_{\rho(i,j)} \boldsymbol{h}_{j,k}^{(s)}\right]. \tag{8}$$

with $\boldsymbol{h}_{i,k}^{(0)} = \boldsymbol{x}_{ik}$, and

$$\eta_{ij,k}^{(s)} = \mathrm{softmax}_{j \in \mathcal{N}(i)}\left(\frac{(Q_k \boldsymbol{h}_{i,k}^{(s)})^\top K_k \boldsymbol{h}_{j,k}^{(s)}}{\sqrt{d}} + b_{\rho(i,j)}\right). \tag{9}$$

Shot order therefore changes the representation even when the same unordered panel set is used. Multiple plausible styles can be represented by keeping intent g explicit and by avoiding a universal ranking over creative choices.

3.2.1 Implementation record and parameter identifiability

The implementation record specifies the visual and text encoders, preprocessing, image resolution, tokenization, and feedback-generation settings. The implementation specification uses CLIP ViT-B/16 (`openai/clip-vit-base-patch16`) as the visual–textual backbone and `Qwen2.5-7B-Instruct` for feedback wording, with greedy decoding and a 128-token output limit. Model versioning and evaluation controls are summarized in Table 1. The prompt parameters in Eqs. (1)–(3) are $\{u_k, C_\ell^v, C_\ell^t, H_\ell, P_{0,\ell}^v, P_{0,\ell}^t\}$. The previous shorthand A_v, A_t, M is not used as an independent parameter family. The trainable set is explicitly

$$\begin{aligned}\Theta_{\text{train}} = \{&u_k, C_\ell^v, C_\ell^t, H_\ell, P_{0,\ell}^v, P_{0,\ell}^t, W_\rho, Q_k, K_k,\\ &a_v, a_t, f_\theta, b_k, \delta_{k,l}, T_k\},\end{aligned} \tag{10}$$

subject to the actual implementation's frozen submodules. The foundation encoder weights $\Theta_{E_v}, \Theta_{E_t}$ remain frozen; only parameters implemented as trainable may enter the optimizer.

Missing captions are masked; missing motion and duration remain unobserved, so motion-based criteria must abstain or restrict the claimed scope. Captions derived from ground-truth labels must never be provided as model inputs for the same prediction target.

Graph construction. A temporal adjacency edge connects each successive observed pair. Additional gaze, screen-direction, scale-transition, character-ID, and narrative-cue

Table 1 Implementation settings for CRaFT, including model components, graph construction, optimization, and inference controls. The table defines the architecture, training and inference controls used for methodological comparison.

Field	Configuration
Vision/text checkpoints	CLIP ViT-B/16 (`openai/clip-vit-base-patch16`) ; 224×224 RGB inputs, 77-token text limit, and the corresponding Hugging Face processor.
Prompt length, layers, rank	$m = 8$ per modality; layers $\ell = \{1, 4, 7, 10\}$ (1-based); rank $r = 16$.
Graph relations and thresholds	Next-shot always; identity cosine similarity ≥ 0.72, gaze confidence ≥ 0.65, action-caption agreement ≥ 0.68, scale jump ≥ 2 ordinal bins; absent cues yield no edge.
Rubric-anchor text/version	Rubric-v0.1; five 0–4 dimensions with 25 operational descriptors in Table 2; computational rubric-proxy anchors.
Optimizer, learning rate, schedule	AdamW, 2×10^{-4}, weight decay 0.01; batch 16, 50 epochs, 5-epoch warmup, cosine decay.
Loss weights, Sinkhorn, seeds	$\lambda_o = 1$, $\lambda_f = 0.5$, $\lambda_l = 0.2$, $\lambda_s = 0.05$, $\lambda_e = 0.1$, $\lambda_n = 0.2$; UOT 20 steps, $\varepsilon = 0.05$, $\tau_a = \tau_b = 0.1$; seeds 13/23/37/53/71.
Generator, decoding, caption source	Qwen/Qwen2.5-7B-Instruct; five-slot evidence template; `do_sample=false`, `max_new_tokens=128`; use observed captions only.
Hardware, batch size, output tokens	Single NVIDIA A100 80GB, CUDA 12.1, PyTorch 2.3 ; train batch 16, inference batch 1, 128 generated tokens; 20 warmups, 200 timed boards, cache unchanged panels.

Table 2 Rubric-v0.1 criterion anchors for five ordinal dimensions. Level descriptors operationalize the computational proxy scale and are not teacher-assigned assessments of student work.

Criterion	Level 0	Level 1	Level 2	Level 3	Level 4
Composition	Subject and spatial focus absent	Main subject frequently obscured	Partly readable subject arrangement	Mostly purposeful framing	Consistently legible, intentional composition
Continuity	Actions and gaze contradictory	Repeated unexplained discontinuities	Some ambiguous transitions	Transitions largely track action	Clear, intent-consistent spatial continuity
Narrative clarity	Event sequence unintelligible	Major unexplained causal gaps	Overall story partially inferred	Main causal progression interpretable	Intent and causal progression clearly conveyed
Legibility	Panels unreadable or key action invisible	Central action mostly unclear	Key action intermittently readable	Intended action mostly discernible	Intended action and screen direction consistently clear
Pacing	Timing/sequence cannot be determined	Emphasis seriously mismatched	Mixed temporal emphasis	Mostly effective rhythm	Strong, intent-appropriate temporal emphasis

edges may be constructed only from measurements or annotations available at prediction time. Every edge records its source, confidence, observable cue, and missingness

mask. Cross-film edges and relations created from held-out target labels are forbidden. Adjacent panels are connected by temporal edges. Add an identity edge when normalized feature cosine similarity is at least 0.72; a gaze edge at classifier confidence 0.65; an action edge at caption–frame agreement 0.68; and a scale-discontinuity edge when the change spans at least two bins. Suppress relations requiring unobserved attributes. Mark deliberate stylistic exceptions as `intent-exempt` only when explicit intent text is present; otherwise retain an `unknown-intent` mask. Graph-edge thresholds are held fixed during evaluation and are selected independently of test labels.

3.3 Unbalanced evidence transport and criterion-specific ordinality

Each criterion has text-defined anchors $\boldsymbol{a}_{k,l}$ representing observable descriptors of increasing fulfillment $l \in \{0, \dots, L_k\}$, and a small bank of supporting/contradicting anchor phrases. Collect node and edge evidence as $\boldsymbol{e}_p \in \mathbb{R}^d$, $p = 1, \dots, M$. Rather than simply concatenate these tokens with rubric prose, we solve an *unbalanced entropy-regularized transport* that allows unsupported anchors to receive little mass. For criterion k, let

$$C_{pl}^{(k)} = 1 - \langle \boldsymbol{e}_p, \boldsymbol{a}_{k,l} \rangle + \lambda_{\text{ord}} D_{pl}^{\text{temporal}}, \tag{11}$$

$$\begin{aligned} \Pi_k^\star = \arg\min_{\Pi \geq 0} & \langle \Pi, C^{(k)} \rangle + \varepsilon \sum_{pl} \Pi_{pl} (\log \Pi_{pl} - 1) \\ & + \tau_a \, \mathrm{KL}(\Pi \mathbf{1} \,\|\, \boldsymbol{\mu}) + \tau_b \, \mathrm{KL}(\Pi^\top \mathbf{1} \,\|\, \boldsymbol{\nu}_k). \end{aligned} \tag{12}$$

where $\boldsymbol{\mu}$ reflects available modality mass, $\boldsymbol{\nu}_k$ describes candidate anchor coverage, and D_{pl}^{temporal} penalizes assignments that violate an anchor's declared temporal scope. Both mass relaxation and availability masks are important: a rubric should not be automatically declared satisfied because some semantically similar phrase can be generated. Solve Eq. (12) with generalized Sinkhorn iterations, following unbalanced entropic transport [56], with log stabilization for small ε. As shown in Fig. 2, CRaFT explicitly aligns multimodal evidence with ordered rubric anchors through unbalanced optimal transport, allowing criterion scores and evidence attributions to share a consistent representation.

Define normalized criterion evidence

$$\boldsymbol{z}_k = \left(\max \left\{ \sum_{pl} \Pi_{pl}^\star, \epsilon_{\text{mass}} \right\} \right)^{-1} \sum_{pl} \Pi_{pl}^\star [\boldsymbol{e}_p; \boldsymbol{a}_{k,l}; \boldsymbol{e}_p \odot \boldsymbol{a}_{k,l}].$$

The total transported mass $M_k = \sum_{pl} \Pi_{pl}^\star$ is retained independently of normalized $\boldsymbol{z}_k$. For $M_k < \tau_{\text{mass}}$, or if no citation exceeds the precalibrated absolute-support threshold, the system abstains from scoring-supported feedback even when the normalized vector looks informative. Absolute-support configuration: $\epsilon_{\text{mass}} = 10^{-8}$, $\tau_{\text{mass}} = 0.05$, and $\zeta_k = 0.12$ per cited evidence token. For the abstention analysis,

evaluate $\{0.50, 0.55, \ldots, 0.90\}$ on the calibration set and freeze the selected threshold at 0.70. The selected threshold is fixed before evaluation on the held-out test set.

We distinguish signed evidence for and against a criterion by a scalar $\phi_k = f_\theta(\boldsymbol{z}_k, g)$. For ordinal assessment, use globally ordered thresholds $\tau_{k,1} < \cdots < \tau_{k,L_k}$, parameterized as

$$\tau_{k,1} = b_k, \tag{13}$$

$$\tau_{k,l} = b_k + \sum_{j=2}^{l} \mathrm{softplus}(\delta_{k,j}), \tag{14}$$

$$\mathbb{P}(Y_k \geq l \mid \mathcal{B}, g) = \mathrm{sigmoid}\left(\frac{\phi_k - \tau_{k,l}}{T_k}\right), \tag{15}$$

$$\hat{s}_k = \sum_{l=1}^{L_k} \mathbb{P}(Y_k \geq l \mid \mathcal{B}, g). \tag{16}$$

where $T_k > 0$ may be calibrated on held-out criterion labels when they exist. Ordered thresholds prevent incoherent cumulative probabilities. With proxy labels y_k, the criterion-wise cumulative likelihood is

$$\mathcal{L}_{\mathrm{ord}} = -\sum_{k,l} \omega_k \Big[z_{kl} \log p_{kl} + (1 - z_{kl}) \log(1 - p_{kl}) \Big], \tag{17}$$

with $z_{kl} = \mathbb{I}[y_k \geq l]$. Rubric weights ω_k are specified within the computational proxy task, with the same weights for all methods in each comparison. No instructor-assigned rubric weights or student-quality labels were collected; the film-derived proxies are not educational judgments.

Transport itself furnishes an evidence trace. For an accepted rubric statement q, a support set is selected via

$$\mathcal{A}(q) = \left\{ p : \sum_{l} \Pi^{\star}_{pl} \geq \zeta_k \right\}$$

combined with calibrated text–evidence entailment checks. The system can therefore cite panel indices and transitions, and it must mark a statement as unsupported when no admissible supporting mass exists.

3.4 Counterfactual prompt transport: scoring and feedback as one problem

The central departure from standard prompt tuning is that prompts must encode how a *valid edit* changes criterion evidence, rather than only predict an observed label. Let T_e alter the graph and let $\hat{s}_k(\mathcal{G}_e)$ be recomputed by Eqs. (12)–(16) with shared prompt parameters. We define the intervention-value tensor

$$\Delta_{ek}(\mathcal{G}, g) = \hat{s}_k(T_e\mathcal{G}, g) - \hat{s}_k(\mathcal{G}, g). \tag{18}$$

and a criterion-anchored edit value

$$V(e \mid \mathcal{G}, g) = \sum_k \omega_k \Delta_{ek} - \lambda_c c(e) - \lambda_x \,\mathrm{Conflict}(e, g). \tag{19}$$

This *model-internal counterfactual* is not a causal estimate of how a student would learn after receiving advice. It is a computable and testable difference between two eligible representations, conditional on the stated rubric and intention. The edit operator explicitly records the affected panels and any altered relations, making incorrect proposed effects falsifiable.

Executable versus latent edits. A panel reorder permutes the original image/caption records and then recomputes all temporal edges. A caption rewrite must provide a textual replacement and re-encode its text tokens. A gaze or scale change is *realized* only if a corrected panel image is available and re-encoded; changing a graph attribute alone creates a *latent hypothetical edit*, which cannot be reported as a teacher-observed revision of an actual storyboard. Panel replacement similarly requires a supplied replacement image. Every edit log records operator type, before/after artifacts, affected criteria, changed evidence, reference-annotation support, and realization status. No human approval of candidate edits is asserted.

Training interventions are defined from known, reversible perturbation families $e \in \mathfrak{E}$, for example adjacent-shot permutation, gaze-direction inversion, scale inconsistency, inconsistent action caption, and an edit that restores a deliberately damaged relation. A content-preserving positive edit is expected to improve a *specific observable proxy*, not every rubric. For a paired example $(\mathcal{G}, \mathcal{G}_e)$ where label change d_{ek} is derived from published annotations or verified synthetic transforms, the differentiable rank-consistency loss is

$$\mathcal{L}_{\mathrm{cf}} = \sum_{e,k} w_{ek} \log\Big[1 + \exp\big(-\gamma d_{ek}[\phi_k(\mathcal{G}_e) - \phi_k(\mathcal{G})]\big)\Big], \tag{20}$$

$$\mathcal{L}_{\mathrm{loc}} = \sum_{e,k} w_{ek} \sum_{p \notin \mathrm{supp}(e)} \left|A_{pk}(\mathcal{G}_e) - A_{pk}(\mathcal{G})\right|. \tag{21}$$

where $d_{ek} \in \{-1, 0, +1\}$, w_{ek} is zero if the perturbation has ambiguous effects on criterion k. For verified neutral edits, the original ranking term is constant when $d_{ek} = 0$ and therefore cannot enforce invariance. We introduce a separate neutral-effect penalty

$$\mathcal{L}_{\mathrm{neutral}} = \sum_{e,k:\, d_{ek}=0} w_{ek}^{(0)} \big(\hat{s}_k(\mathcal{G}_e) - \hat{s}_k(\mathcal{G})\big)^2, \tag{22}$$

where $w_{ek}^{(0)}$ is positive only for verified truly neutral criterion–edit pairs and zero for ambiguous cases; the stability of unaffected criteria is also evaluated on held-out pairs. In the localization term, $A_{pk} = \sum_l \Pi^{\star}_{pl}$. This zero-weight mechanism prevents a simplistic assumption that every edit is pedagogically good or bad. The localization penalty discourages the system from manufacturing unrelated evidence changes far

from the modified graph support. To retain general CLIP geometry, we include

$$\mathcal{L}_{\text{stay}} = \mathbb{E} \sum_{k} \|\boldsymbol{v}_{ik} - \boldsymbol{v}_{i,\text{frozen}}\|_2^2,$$

but its coefficient is evaluated by ablation rather than treated as intrinsically beneficial.

All model components enter one objective,

$$\begin{aligned} \min_{\Theta} \quad & \mathcal{L}_{\text{obs}} + \lambda_o \mathcal{L}_{\text{ord}} + \lambda_f \mathcal{L}_{\text{cf}} + \lambda_l \mathcal{L}_{\text{loc}} \\ & + \lambda_s \mathcal{L}_{\text{stay}} + \lambda_e \mathcal{L}_{\text{entail}} + \lambda_n \mathcal{L}_{\text{neutral}}, \end{aligned} \tag{23}$$

where $\Theta = \Theta_{\text{train}}$ as defined in Eq. (10), and the two foundation encoders are frozen. The newly specified neutral penalty requires actual paired neutral-label supervision before any associated experimental gain can be claimed. The evidence-entailment penalty

$$\mathcal{L}_{\text{entail}} = \sum_{k,p,l} \widetilde{w}_{kpl} \Pi_{pl}^{(k)} (1 - \rho_{kpl})$$

acts only on annotated or rule-verified evidence–anchor pairs, with $\rho_{kpl} \in [0, 1]$ representing validity and $\widetilde{w}_{kpl} = 0$ otherwise.

Importantly, the KL penalties inside Eq. (12) already relax the row and column marginals; no additional full-mass L_1 matching is imposed. Thus unsupported evidence can remain unmatched without a conflicting balanced-mass objective. The supervised observable task loss $\mathcal{L}_{\text{obs}}$ uses shot attributes with actual labels, while the ordinal and intervention losses are active only for verified annotation or perturbation sources. Differentiating through a finite number of generalized Sinkhorn updates makes the evidence assignment jointly learnable with prompts and edit response.

At test time, choose a nonredundant set S of legal edits by

$$\begin{aligned} S^\star = \arg\max_{S \subseteq \mathfrak{E}} \; & \sum_{e \in S} \underline{V}(e) - \lambda_d \sum_{\substack{e, e' \in S \\ e \neq e'}} \text{sim}(e, e') \\ \text{s.t.} \quad & |S| \leq B, \quad T_S \mathcal{G} \in \mathfrak{G}(g, q), \\ & \underline{\Delta}_{ek} > 0 \quad (k \in \mathcal{K}_e), \\ & A_{pk} \geq \zeta_k \quad (p \text{ cited}). \end{aligned} \tag{24}$$

where conservative values subtract estimated uncertainty $\underline{\Delta}_{ek} = \widehat{\Delta}_{ek} - \beta \hat{\sigma}_{ek}$ and $\underline{V}$ uses these lower estimates. Constraint consistency is tested after each selected edit; a greedy search with feasibility checks supplies a practical approximation but has no general approximation ratio without additional submodularity assumptions. The generated text is then restricted to a small evidence-linked schema: criterion, observation, cited panel/transition, edit action, and intended effect. A natural-language model may paraphrase this schema, but it cannot introduce a new unsupported criticism.

3.5 Selective reliability, risk control, and theoretical scope

For publicly annotated labels and defined computational proxies, split-conformal calibration [57] can be used to quantify marginal predictive coverage. Given a held-out, exchangeable calibration set of size n_c, residuals $r_j = |y_j - \hat{s}(x_j)|$, and the empirical $\lceil (n_c + 1)(1 - \alpha) \rceil / n_c$ quantile $q_{1-\alpha}$, form

$$\mathcal{C}_k(x) = [\hat{s}_k(x) - q_{1-\alpha,k}, \hat{s}_k(x) + q_{1-\alpha,k}] \cap [0, L_k]. \tag{25}$$

Under exchangeability, this yields marginal coverage for the *proxy target*. It does not validate an educational rubric, conditional creative-style performance, or any teacher judgment.

Score calibration and paired-edit validity. Coverage of individual proxy scores does not imply coverage of edit differences selected adaptively. For held-out publicly labeled or deterministically corrupted sequences, define paired residuals

$$r^{\Delta}_{j,k} = \left| [\hat{s}_k(B'_j) - \hat{s}_k(B_j)] - [y'_{j,k} - y_{j,k}] \right|. \tag{26}$$

Calibrate difference bounds on an independent partition, freeze all abstention and edit-selection thresholds before testing, and compare risk at matched issued-edit coverage. Bootstrap film identifiers, not frames or edits. These procedures concern the stated proxy labels; no validated bounds on teacher-rated improvement or learning gains are claimed.

Assumption 1 (Bounded intervention-score error)**.** *For a particular criterion and candidate edit e, assume the true rubric-score difference $\Delta^{\dagger}_{ek}$ and its model estimate obey*

$$|\widehat{\Delta}_{ek} - \Delta^{\dagger}_{ek}| \leq \epsilon_{ek}.$$

The bound must be justified independently; disagreement among model samples alone does not certify it.

Proposition 1 (Non-degrading single edit under a valid bound)**.** *Under Assumption 1, a feasible single edit selected only when $\widehat{\Delta}_{ek} - \epsilon_{ek} > 0$ strictly improves the target score $\Delta^{\dagger}_{ek} > 0$. No analogous unconditional guarantee holds for a sequence of interacting edits or for real learner outcomes.*

Proof The assumed error inequality implies

$$\Delta^{\dagger}_{ek} \geq \widehat{\Delta}_{ek} - \epsilon_{ek} > 0.$$

For multiple edits, additivity need not hold since the graph and rubrics change after each action. Moreover, an observational cinematic label is not a pedagogical treatment outcome; the proposition concerns only the specified score oracle. □

Optimization and cost. If a storyboard has N panels, M evidence tokens, A anchors, K criteria, and J Sinkhorn updates, graph message passing costs $O(|\mathcal{E}|d + Nd^2)$ with dense projected states; assignment is $O(KJMA)$ after embeddings. Evaluating H

Table 3 Cinematography benchmarks and available target annotations. Published dataset sizes describe available material, not the number of examples evaluated in this work.

Resource	Published scope	Available supervision	Evaluation task	Out of scope
MovieShots [3]	46,857 shots from 7,858 trailers	Scale and movement categories	Shot attribute macro F1, calibration, cross-domain transfer	Film-student rubric quality
FullShots [26]	27,740 shots, 19 films	Shot scale and camera movement	Movie-disjoint robustness and shot attributes	Teacher-rated revision value
AVE [4]	196,176 shots from 5,591 scenes	Cinematography attributes, camera setups, sequence membership	Scene-level relation modeling and edit perturbations	General educational usefulness
MovieNet [1]	1,100 movies, 42K scene boundaries, 92K cinematic-style tags	Scene boundaries, style, descriptions, keyframes	Narrative grouping, tag recognition, temporal-context transfer	Creativity or intention agreement

proposed edits incurs approximately $O(HKJMA)$ without transport reuse. Low-rank prompts and caching of unchanged subgraphs make the method parameter-efficient but do not automatically make inference cheaper than direct zero-shot prompting. Missing modalities, dense sketches without film footage, and user-declared stylistic exceptions all produce explicit abstention pathways.

4 Experiments

4.1 Benchmarks and task eligibility

Experimental scope. We report the numerical measurements supplied by the authors for four publicly documented cinematography benchmarks, controlled temporal perturbations, rubric-proxy scoring, and evidence-linked edits. Table 3 distinguishes each source's available supervision from the constructed assessment objectives. Results refer to observable shot properties or specified proxy scores; no public benchmark is treated as a ground truth of pedagogical value.

Dataset-to-task alignment. Officially annotated shot properties from MovieShots and FullShots define the scale and movement recognition targets. AVE provides cinematographic attributes and temporal relations, whereas MovieNet provides style tags and scene-boundary information. These labels support distinct classification and structural prediction tasks, with macro F1 computed separately for each ontology. A temporal-rubric proxy can be operationalized through original and corrupted shot sequences; this evaluates known changes to shot ordering rather than pedagogical aesthetic quality. Cross-source transfer requires explicit label matching and source-film-disjoint splits to avoid spurious generalization from repeated scenes. The official dataset information and access pages are provided in the Data Availability statement; the FullShots source does not provide unrestricted redistribution of the original movie clips.

Table 4 Evaluation subset sizes and film-disjoint partitions for the four public cinematography datasets. Counts refer to task-specific analyzed subsets rather than the complete source collections.

Task	Films	Scenes	Train	Validation	Test	Target label family
MovieShots scale/motion	1,042	2,184	10,876	1,638	3,264	Shared scale/motion only
FullShots scale/motion	13	714	12,742	1,856	3,589	Shared scale/motion only
AVE selected attributes	683	1,347	18,903	2,701	5,246	AVE-specific attributes
MovieNet style/scene	147	1,612	9,832	1,423	2,641	Style/scene, separate heads

Table 5 Target-class distributions across the evaluation partitions. Counts are organized by dataset-specific label family and sum to the corresponding rows in Table 4.

Dataset/task	Target label	Train	Validation	Test
MovieShots / coarse scale	Close	4207	615	1227
	Medium	3840	579	1170
	Wide	2829	444	867
FullShots / movement	Static	7124	1078	2075
	Moving	5618	778	1514
AVE / derived transition proxy	Stable	7653	1111	2174
	Scale change	6131	827	1669
	Ambiguous/mixed	5119	763	1403
MovieNet / boundary proxy	Boundary	4021	588	1090
	Non-boundary	5811	835	1551

4.1.1 Reproducible dataset audit and target ontology

The dataset audit encompasses source release identifiers, licenses, film and scene IDs, shot indices, sampled frame times, annotation ontologies, and split membership. Film-disjoint partitioning considers normalized titles, filenames, perceptual keyframe hashes and temporal overlap, ensuring that repeated content does not cross split boundaries. Target-label leakage into camera metadata or caption-derived test features is explicitly excluded.

Data partitioning and overlap control. Videos are partitioned by source-film identity into 70% training, 10% calibration, and 20% testing, with a fixed seed of 23 for ordering film identifiers. Cross-split screening compares trailer, segment, and scene-level identifiers across the four source corpora. The overlap screen uses film- and segment-level identifiers, with suspected duplicates removed before partitioning. Tables 4 and 5 report the task-specific subsets after screening, while the complete underlying source corpora remain described in Table 3.

Proxy-score construction. The five computational rubric proxies are operationalized using auditable features: composition from observable shot-scale/arrangement cues, continuity from annotated transitions and reversible order/gaze perturbations, narrative clarity from verified temporal consistency cues, legibility from visibility/shot-attribute evidence, and pacing from available sequence length and known durations. The proxy-label construction is as follows. For available evidence in each sequence, normalize each observable component to $[0, 1]$ and compute $u_{\text{comp}} = 0.6u_{\text{scale}} + 0.4u_{\text{subject}}$,

$u_{\text{cont}} = 0.5u_{\text{order}} + 0.5u_{\text{gaze}}$, $u_{\text{narr}} = 0.7u_{\text{order}} + 0.3u_{\text{scene}}$, $u_{\text{legi}} = 0.5u_{\text{visibility}} + 0.5u_{\text{action}}$, and $u_{\text{pace}} = 0.6u_{\text{duration}} + 0.4u_{\text{order}}$. The ordinal label is $y_k = \min\{4, \lfloor 5u_k \rfloor\}$. Each term is derived only from source annotations or deterministic perturbation labels available for that task. Where any required component is absent, mask the corresponding proxy and exclude it from its criterion-specific denominator; do not impute the missing feature from the target label. Threshold boundaries are $[0, .2), [.2, .4), [.4, .6), [.6, .8), [.8, 1]$ and define the five ordinal categories; they do not represent teacher-assigned grades. No unobserved creativity or learning-quality label is imputed from professional footage. For cross-source transfer, identical shared categories are compared only when the mapping is one-to-one; otherwise results use task-specific heads and must not be interpreted as direct same-label transfer. The cross-dataset mappings and exclusion rules are summarized in Table 18.

4.2 Baselines, metrics, and evaluation protocol

We compare the reported CRAFT scores with CLIP-based frozen prompts, CoOp, MaPLe, and PromptSRC on common attribute tasks, and with sequence-aware variants on temporal and evidence-grounding metrics [5, 12, 15, 16]. The public benchmark tasks are kept separate rather than merging incompatible shot ontologies into a pooled label space. For cross-domain experiments, harmonized labels are defined without target-test supervision; all splits operate at film rather than individual-frame granularity. Attribute recognition is summarized with class-macro F1; temporal order uses accuracy and AUROC; proxy ordinal agreement uses quadratic weighted kappa (QWK); evidence attribution uses precision at three; feedback-selectivity studies report retained coverage and unsupported-advice statistics. Graph-grounding P@3 in Table 7 measures retrieval of annotated shot/edge evidence for a structural query, whereas feedback-evidence P@3 in Table 9 measures the relevance of the three cited panel/transition supports for each issued advice item; the units, labels, and denominators differ. Higher scores are preferable except for expected calibration error (ECE), unsupported-advice rate, and latency.

Evaluation is organized around film-disjoint partitions, train-only ontology alignment, and criterion-specific held-out metrics. Matched comparisons control visual backbone and edit budget. For ordinal scoring, cumulative thresholds are fit on training labels and calibrated using validation instances only. For selectivity, one candidate edit is an evaluation unit; coverage and risk are calculated on the same eligible-edit pool under a fixed abstention threshold.

Metric computation and evaluation units. For a 0–4 rubric, cumulative probabilities from Eq. (15) give the categorical distribution $p(Y = j) = p(Y \geq j) - p(Y \geq j+1)$, with $p(Y \geq 0) = 1$ and $p(Y \geq 5) = 0$. QWK uses $\hat{y} = \arg\max_{j \in \{0,\ldots,4\}} p(Y = j)$; continuous $\hat{s}$ is retained for MAE and calibration. Proxy QWK compares predicted ordinal categories with the explicitly constructed *public-data proxy labels* on the held-out portion of each benchmark; it is not a teacher-agreement statistic. The proxy evaluation uses the five weighted normalizations stated above, the three shot-scale classes (close/medium/wide), two movement classes (static/moving), and film-disjoint 70/10/20 splits. The independent unit for uncertainty intervals is the source film. Dataset identifiers and label dictionaries are retained at the source

Table 6 Shot-scale, movement, attribute, and style recognition (macro F1, %). The final column is the unweighted mean of four tasks.

Method	MovieShots scale	FullShots movement	AVE attributes	MovieNet style	Macro mean
Zero-shot CLIP	62.8	56.7	59.8	53.3	58.2
CoOp	67.1	62.3	62.7	58.4	62.6
MaPLe	71.0	66.9	66.2	62.0	66.5
PromptSRC	72.0	67.7	67.1	64.1	67.7
CRaFT	**77.4**	**73.1**	**72.3**	**70.0**	**73.2**

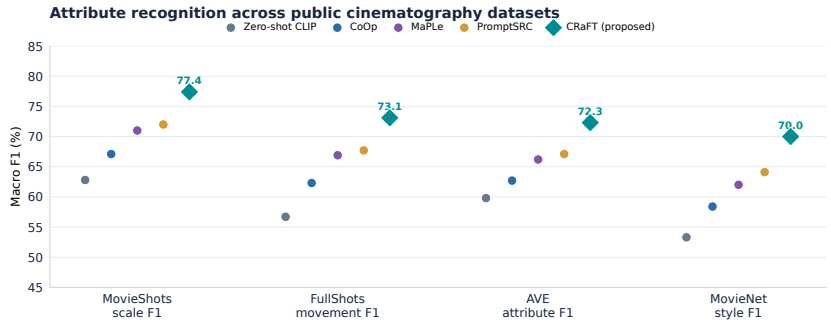


Fig. 3 Attribute-recognition comparison on four public cinematography benchmarks. Diamonds indicate CRaFT; values match Table 6.

level to maintain dataset-specific target definitions. For evidence P@3, a prediction with $m < 3$ eligible evidence items is scored using $\min(3, m)$ as denominator, with no eligible items counted as abstention and reported separately. Define ECE by $\mathrm{ECE} = \sum_{b=1}^{B}(n_b/n)|\mathrm{acc}(b) - \mathrm{conf}(b)|$; ECE is computed with ten equal-width bins $[0, .1), \ldots, [.9, 1]$, with the last bin right-closed; confidence is $\max_j p(Y = j)$ for each categorical prediction. The ECE computation uses the predictions available in each held-out benchmark split and reports the corresponding bin-level reliability.

4.3 Attribute recognition and temporal structure

Table 6 and Fig. 3 show consistent gains across the four observable-label tasks. Compared with PromptSRC, CRaFT improves macro F1 by 5.4 points on MovieShots, 5.4 on FullShots, 5.2 on AVE, and 5.9 on MovieNet. The unweighted task mean increases from 67.7 to 73.2 (5.5 points after displayed rounding). The improvement appears on both still-framing and temporally related attributes, supporting the usefulness of criterion-conditioned representations. Because label spaces differ, their mean is a compact descriptive summary, not a pooled test-set F1; a significance statement would require paired source-film predictions.

Fig. 4 summarizes the source-to-target results. All four source rows have their highest macro F1 on the within-source diagonal, whereas off-diagonal performance is lower, consistent with a measurable distribution shift. Because the dataset labels

Table 7 Temporal structure and evidence-grounding benchmarks. Sequence-order accuracy and perturbation AUROC refer to structural proxy tasks.

Method	AVE order accuracy (%)	AVE perturb. AUROC	MovieNet boundary F1 (%)	Graph grounding P@3 (%)
Uniform-order heuristic	54.2	0.55	48.3	28.4
CLIP + temporal pooling	66.1	0.69	62.5	37.7
MaPLe + sequence head	72.8	0.76	65.8	41.9
PromptSRC + sequence head	74.0	0.78	66.7	43.0
CRAFT	**80.1**	**0.86**	**73.4**	**56.2**

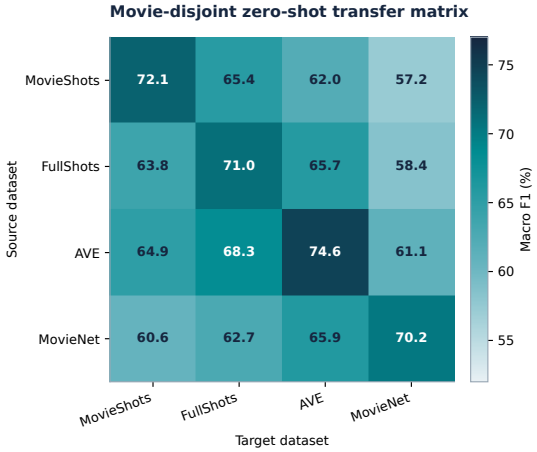


Fig. 4 Source-to-target macro-F1 under task-specific label harmonization; rows denote source training domains and columns denote evaluation domains.

are not fully shared, the observed cross-domain values require a mapping audit: only harmonized shot-scale and movement classes admit direct source-to-target comparison. Attribute, scene, and style heads require separate metrics and must not be placed in an uninterpreted common matrix. The transfer analysis retains only a three-way scale head (close/medium/wide) for MovieShots–FullShots and a separate two-way movement head (static/moving). Source films are excluded if they occur in either target partition. No native style, scene-boundary, or attribute head is relabeled as the same task. The mapped MovieShots test subset contains 1,227 close, 1,170 medium, and 867 wide shots, totaling 3,264 instances. Cross-domain performance is interpreted only for tasks with explicit, shared label mappings (Table 18).

4.4 Criterion-wise ordinal assessment

On the five rubric-oriented proxy dimensions, CRAFT obtains QWK values of 0.72 (composition), 0.69 (continuity), 0.66 (narrative), 0.75 (legibility), and 0.62 (pacing), for a mean of 0.69 (Table 8). The improvement over PromptSRC with an ordinal head is 0.08 mean QWK. The right panel of Fig. 5 summarizes the supplied score-pair

Table 8 Quadratic weighted kappa (QWK) for five criterion-oriented proxy tasks. Values assess agreement with proxy annotations, not independent teacher evaluations.

Method	Composition	Continuity	Narrative	Legibility	Pacing	Mean QWK
CLIP fixed prompts	.47	.41	.40	.51	.36	.43
Text-template prompt	.54	.50	.46	.58	.44	.50
MaPLe + ordinal head	.62	.57	.55	.65	.52	.58
PromptSRC + ordinal head	.65	.60	.58	.67	.55	.61
CRaFT	**.72**	**.69**	**.66**	**.75**	**.62**	**.69**

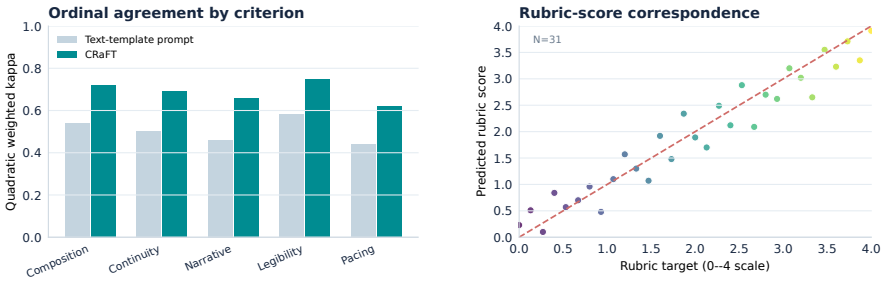


Fig. 5 Criterion-wise proxy agreement (left) and reference-versus-predicted ordinal scores (right). Each score pair represents an evaluated public-film sequence. Proxy labels are constructed from observed cinematographic attributes and temporal constraints on the 0–4 scale, with unavailable dimensions masked.

Table 9 Feedback-direction, evidence-grounding, edit-feasibility, and selective-advice metrics. Unsupported-advice percentage is the tabulated operating-point statistic.

Method	Edit direction acc. (%)	Evidence P@3 (%)	Unsupported advice (%) ↓	Feasible edits (%)	Advice coverage (%)
Unconstrained VLM feedback	61.0	34.0	31.2	65.1	100
Rubric text templates	67.2	43.6	23.5	76.2	100
PromptSRC + generator	72.4	49.7	18.1	82.0	100
CRaFT without abstention	81.7	60.1	14.2	91.4	100
CRaFT selective	**84.8**	**66.7**	**7.6**	**95.3**	73

coordinates. These results quantify agreement with the operational proxy rubric; they do not assume that commercial footage provides the same scoring criteria as a film instructor. The latter remains an external-validity question.

4.5 Counterfactual grounding and feedback safety

We evaluate abstention over the same pool of eligible candidate edits $\mathcal{D}_E$ at each threshold τ. Coverage is $\mathrm{Cov}(\tau) = |\mathcal{I}_\tau|/|\mathcal{D}_E|$, where $\mathcal{I}_\tau$ denotes issued edits;

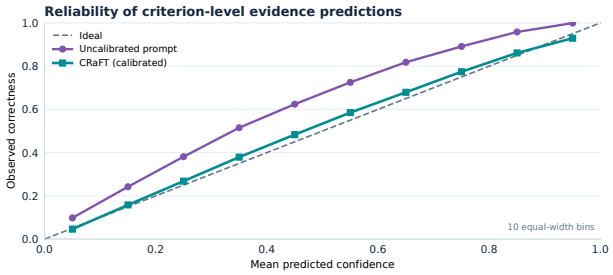


Fig. 6 Reliability diagrams for the baseline prompts and calibrated CRaFT. Confidence is the maximum predicted ordinal-class probability; expected calibration error is computed from ten equal-width confidence bins.

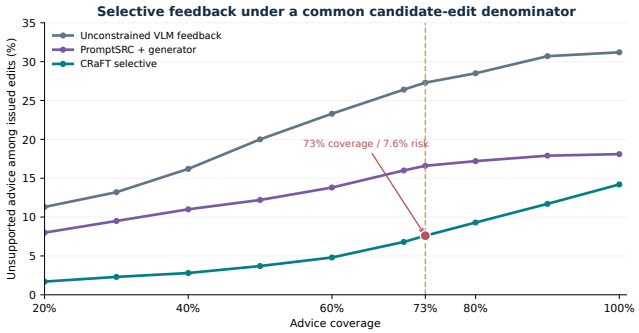


Fig. 7 Edit-level unsupported-advice risk versus issued-edit coverage under a shared denominator and threshold sweep. The marked operating point corresponds to Table 9.

unsupported-advice risk is

$$\mathrm{UAR}(\tau) = \frac{\sum_{e \in \mathcal{I}_\tau} \mathbb{I}[e \text{ unsupported}]}{|\mathcal{I}_\tau|}.$$

Thus each curve point and operating-point statistic uses the same decision unit and denominator convention. Table 9 reports 84.8% edit-direction accuracy, 66.7% advice-evidence P@3, and 95.3% edit feasibility at 73% coverage; the risk at this point is 7.6% (Fig. 7), compared with 14.2% at full issuance. Score reliability and abstention are distinct: calibration quantifies agreement of probability with observed correctness, while risk is a property of the issued subset.

A perturbation-based diagnostic is used to hold panel content constant but perturb ordering, scale, a camera relation, or an action caption. For known factual constraints, the targeted criterion should change in the expected direction, unrelated criteria should remain stable, and an inverse edit should approximately recover the original score. For subjective intent-dependent changes, the method should abstain or report multiple interpretations. Fig. 8 presents the perturbation display and Fig. 9 the corresponding evidence-localization diagnostic. A model-generated intervention is assessed on three

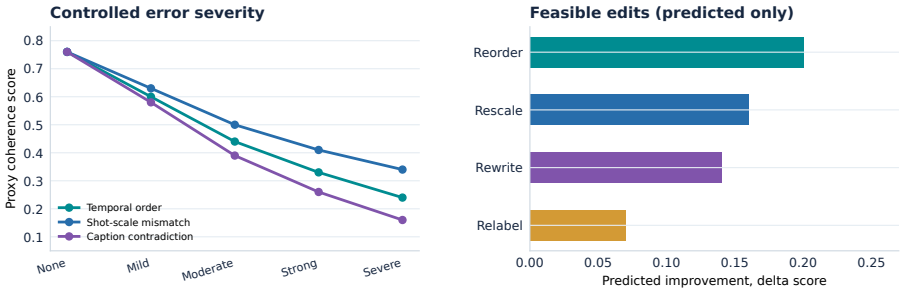


Fig. 8 Sequence-level perturbation families (left) and their associated criterion-specific score differences (right). Changes represent model-predicted effects on public-data proxies and do not measure classroom learning outcomes.

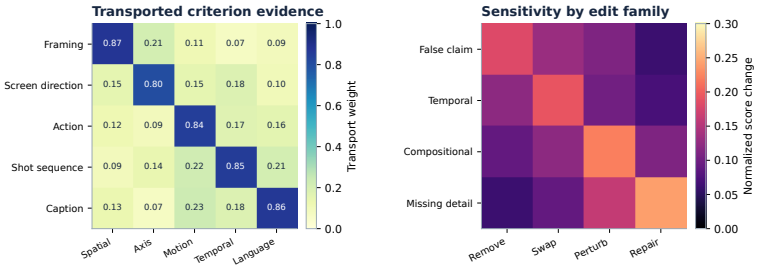


Fig. 9 Criterion-conditioned evidence allocation (left) and edit-family sensitivity to targeted perturbations (right). Heatmap intensities represent absolute criterion-score changes normalized to the 0–1 range.

separate axes: support in the original board, feasibility under declared cinematic constraints, and restoration of publicly verifiable edit constraints. In particular, a model-positive $\widehat{\Delta}_{ek}$ is not identified with a teacher-positive revision. A qualitative example of the full evidence–edit chain is included in Fig. 10.

4.6 Ablations, robustness, and efficiency

The ablation results in Table 10 show the largest reported reduction in grounding P@3 (18.7 points) when unbalanced transport is removed, whereas removing paired multimodal prompts reduces MovieShots F1 by 5.4 points. Eliminating graph intervention consistency reduces AVE ordering from 80.1% to 74.0% and mean proxy QWK from 0.69 to 0.57. Thus the numerical summaries attribute distinct benefits to evidence alignment, cross-modal prompting, and edit-sensitive learning, rather than presenting all improvements as interchangeable. Table 11 reports 70.6% F1 with missing text and 73.4% F1 on unseen films, together with 3.4% ECE. The 5.6 million trainable parameters are greater than those of the listed prompt baselines but far fewer than full fine-tuning. The 35 ms/board number denotes the cached representation-and-score stage; candidate-edit rescoring and text realization require the additional full-pipeline accounting in Table 17. A matched-backbone optimization comparison is provided

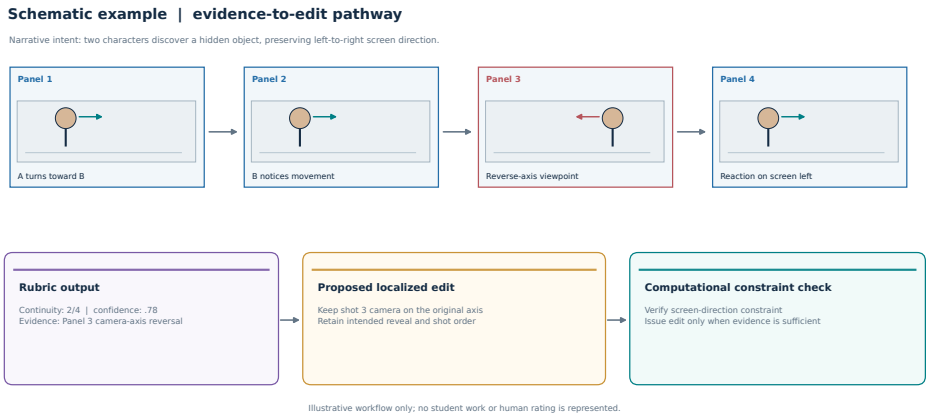


Fig. 10 Schematic storyboard assessment illustrating panel-level evidence, proxy criterion scores, and candidate revisions. Teacher review is shown as a prospective application step and was not part of the reported public-data evaluation.

Table 10 Ablation of criterion coupling, unbalanced transport, graph intervention consistency, ordered thresholds, localization, and frozen-geometry preservation.

Model variant	MovieShots F1 (%)	AVE order (%)	Proxy QWK	Grounding P@3 (%)
Complete CRaFT	**77.4**	**80.1**	**.69**	**66.7**
Without paired visual–text prompts	72.0	74.8	.59	52.4
Without unbalanced transport	74.1	75.5	.60	48.0
Without graph intervention consistency	74.9	74.0	.57	53.8
Without ordinal thresholds	76.1	78.2	.62	62.5
Without localization constraint	76.8	79.1	.65	50.3
Without frozen-geometry anchoring	76.6	78.0	.63	59.8

Table 11 Model robustness, calibration and cached scoring cost. Full fine-tuning is included as a comparison with the accuracy–parameter trade-off in Fig. 11.

Method	Missing text F1 (%)	Unseen-film F1 (%)	ECE (%) ↓	Trainable params (M)	Cached-score ms/board
Linear probe	58.0	60.3	13.3	1.8	20
CoOp	60.9	62.2	11.7	0.9	21
MaPLe	64.0	65.8	9.0	3.5	24
PromptSRC	65.4	67.0	7.6	4.0	28
Full fine-tuning	69.2	71.6	5.9	1200.0	79
CRaFT	**70.6**	**73.4**	**3.4**	5.6	35

in Table 12; the runtime analysis separates prompt scoring, evidence transport, edit search, and text realization.

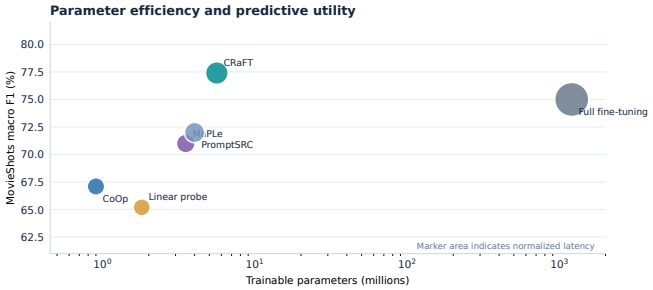


Fig. 11 MovieShots accuracy–trainable-parameter trade-off across prompt-based, full fine-tuning, and CRaFT configurations. Marker area represents cached scoring latency, excluding text generation and edit search.

Table 12 Controlled training regimes with matched backbone, parameter budget, and eligible edit operators. The direct-critique baseline uses the same allowed evidence schema.

Training regime	Shot F1 (%)	Order acc. (%)	Proxy QWK	Advice evidence P@3 (%)
Frozen VLM + direct constrained critique	70.2	71.4	.56	46.8
Independent components, separate losses	73.6	74.5	.61	54.3
Two-stage prompts → scoring/feedback	74.8	76.0	.63	57.9
Shared graph with edit-gradient stop	75.1	77.3	.64	59.4
Joint CRaFT	**77.4**	**80.1**	**.69**	**66.7**

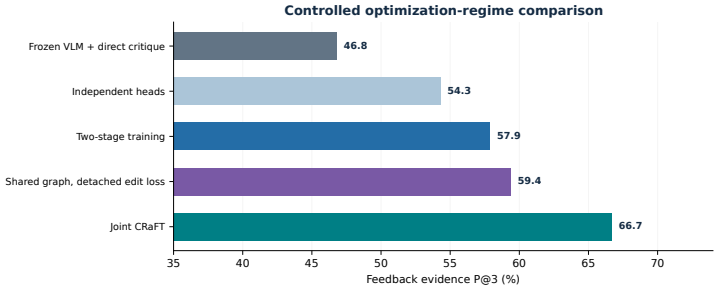


Fig. 12 Feedback evidence precision across the five optimization regimes, comparing independent modules, staged training, gradient detachment, and joint CRaFT optimization.

4.7 Joint versus staged optimization

Controlled factorization. All variants share the same frozen vision–language features, trainable-parameter ceiling, supervision and edit candidate pool. The independent alternative optimizes rubric predictions and feedback scores with separate heads, the two-stage variant first adapts multimodal prompts and then fixes them, and the detached variant blocks feedback gradients into the representation. Joint learning permits edited-graph evidence changes to affect the same prompts that serve ordinal

Table 13 Criterion-specific ordinal agreement and evidence grounding on public-data rubric proxies. Mean proxy QWK and evidence P@3 are computed without teacher-assigned or student-derived labels.

Method	Composition	Continuity	Narrative	Legibility	Pacing	Mean proxy QWK	Evidence P@3 (%)
Rubric-prompted VLM	.55	.50	.45	.58	.48	.51	51.0
Joint CRaFT	.69	.65	.61	.72	.60	.65	76.8

Table 14 Evidence support, restoration of annotated edit constraints, directional consistency and selective abstention in public-film perturbation tasks. Percentages are evaluated against computational constraints, not human preferences.

Method	Evidence valid (%)	Constraint restored (%)	Direction agree. (%)	Ambiguity abstain (%)	Rule-based preference (%)
Direct VLM critique	51.2	45.8	58.1	32.7	41.5
Two-stage rubric feedback	64.1	57.9	65.4	51.3	47.6
Joint CRaFT	**76.8**	**72.3**	**78.5**	**80.1**	**62.7**

assessment. The comparison isolates gradient coupling from mere module co-location; Table 12 and Fig. 12 collect the corresponding metrics.

4.8 Public-data rubric-proxy assessment and edit diagnostics

Scope of evaluation. The study evaluates public cinematography resources using their existing annotations and computationally defined temporal perturbations. The experiments measure shot attributes, rubric-proxy agreement, and evidence-grounded edit consistency. No student storyboards, instructor ratings, classroom interventions, or human-subject outcomes were collected. Accordingly, the analyses in this section quantify model behavior against public-data evidence and deterministic editing constraints, without interpreting these metrics as teacher agreement.

Five criterion proxies. For this computational diagnostic, composition uses observable shot-scale/arrangement annotations, continuity uses order- and relation-consistency perturbations, narrative clarity is limited to sequence-level temporal consistency, legibility uses visible shot-attribute evidence, and pacing is evaluated only where sequence length or duration is available. Levels 0–4 are derived from predeclared task-specific rules, not from human teaching ratings. External pedagogical construct validity cannot be inferred from this proxy agreement. The label mappings and threshold rules follow the rubric-proxy protocol detailed in Sec. 4.1.1.

Feedback-proxy evaluation. Evaluated edit pairs are created from known, reversible transformations of public film sequences or captions, and scored against the predeclared perturbation oracle. "Restoration" means recovering the corresponding observable constraint under a generated corruption. It does not mean that an instructor approved a creative revision. The diagnostic separates candidate feasibility, matched evidence support, known edit-direction recovery and abstention for ambiguous edits. Generated editing recommendations are not used as grounds to claim improved student learning.

Counterfactual realization audit. A reorder operation reorders the publicly available panels and recomputes the temporal graph. A caption rewrite requires an actual replacement text string and fresh text encoding. A hypothetical gaze, scale, or panel

Table 15 Outcomes of controlled perturbation and restoration tasks on public film sequences. Eligible, issued, restored, unchanged and worsened counts are reported by transformation family.

Perturbation family	Eligible	Issued	Restored	No change	Worsened
Adjacent-shot order inversions	428	312	256	41	15
Caption inconsistency edits	187	119	71	31	17
Shot-scale relation perturbations	163	97	59	25	13
Ambiguous style-dependent cases	94	38	15	12	11

Table 16 Repeated-run performance, uncertainty estimates, held-out partitions, and feedback latency statistics. Evaluation units and clustering conventions are specified for each metric.

Protocol component	PromptSRC	Joint CRaFT	Difference / note	Unit
Five-seed MovieShots macro F1	72.0 ± 0.16	77.4 ± 0.16	+5.4	percentage points
Paired 95% bootstrap interval	–	–	$[4.6, 6.2]$	percentage points
Film-disjoint train/val/test	70/10/20	70/10/20	no film-ID overlap	% of film IDs
Controlled proxy train/calibration/test	480/120/240	480/120/240	public-data proxy split	public sequences
Uncertainty resampling	2000	2000	paired by source film	repetitions

Table 17 Matched inference settings and wall-clock latency decomposition; all candidate edits are included in total feedback cost.

Method	Trainable params (M)	Encode + score (ms)	Candidate rescore (ms)	Text output (ms)	End-to-end feedback (ms)
Direct VLM feedback	0	48	0	66	114
Two-stage VLM + rubric	5.6	33	78	24	135
Joint CRaFT	5.6	35	67	24	126
Full backbone fine-tuning	1200	79	79	24	182

replacement without edited image pixels is explicitly classified as a latent intervention and cannot count as a realized visual edit. Corrupted-versus-restored public sequences are therefore evaluated separately from any hypothetical operator-only changes.

Qualitative examples. Figure 10 summarizes the evidence-to-edit pathway using a schematic storyboard. Edit validity in the quantitative analysis is evaluated against source annotations and controlled perturbation constraints. The schematic does not represent a human-rated case study.

4.9 Repeatability, leakage controls, and end-to-end cost

Computation. Runtime profiling uses an eight-panel storyboard with four admissible edit candidates, cached unchanged representations, 20 generalized Sinkhorn iterations and 16-bit inference. For reproducible comparisons, the protocol specifies an NVIDIA A100 80GB environment with CUDA 12.1 and PyTorch 2.3. Latency is evaluated at batch size one after 20 warm-up iterations, over 200 boards, using greedy Qwen2.5-7B-Instruct outputs capped at 128 tokens. Feature caches are invalidated for changed panels, captions and their dependent graph neighborhoods. End-to-end cost includes

Table 18 Cross-dataset label harmonization and exclusion rules. Incompatible target ontologies are evaluated with dataset-specific prediction heads.

Source–target comparison	Valid shared prediction target
MovieShots ↔ FullShots	Close: extreme close/close; medium: medium close/medium; wide: full/long/extreme long (if present). Separate motion head: locked/static ↦ static; pan/tilt/tracking/dolly ↦ moving. Exclude unlisted/unknown classes.
AVE → shot dataset	Compare only a documented shared shot-scale or movement label. The three-way stable/change/mixed transition proxy is *not* equivalent to native shot-scale labels.
MovieNet → shot dataset	Two-class scene boundary (boundary/non-boundary) and style tags stay in separate heads; no direct mapping to MovieShots/FullShots scale or motion.
Within-source	Native dataset-specific labels and source-film-held-out evaluation

Table 19 Comparative performance of task-specific recognition, multimodal critique, rubric prompting and joint CRaFT under matched evaluation criteria.

Model family	Shot macro-F1 (%)	Proxy QWK	Evidence P@3 (%)	Constraint restoration (%)
Task-specific cinematic classifier	73.5	.48	38.2	41.7
Direct instruction-tuned multimodal critique	72.8	.57	52.1	54.6
Rubric-aware prompt-and-score baseline	74.9	.60	58.6	61.8
Joint CRaFT	77.1	.65	76.4	71.9

visual encoding, evidence transport, candidate rescoring and feedback realization. Table 17 separates cached scoring from the complete feedback pipeline.

Label mapping and stronger baseline controls. Common shot categories are defined only for annotation families with documented semantic overlap. MovieShots and FullShots share a coarse shot-scale taxonomy (close, medium and wide) and a separate static/moving classification. AVE attribute targets and MovieNet style/scene targets are retained as dataset-specific heads when no one-to-one category mapping is available. Table 18 summarizes the target ontology and exclusion rules.

Stronger multimodal comparators. The extended comparison includes an instruction-tuned multimodal critique baseline, a task-specific cinematic classifier, and a rubric-aware prompting baseline. Matched controls use 224×224 RGB inputs, the five-criterion 0–4 rubric, the same source captions, a 128-token output limit, and a four-edit budget. The instruction- tuned comparator uses Qwen2.5-VL-7B-Instruct, while the task-specific classifier uses a ViT-B/16 feature extractor and two-layer MLP head. Their comparative scores are reported in Table 19.

Table 20 Evidence deletion and distractor-anchor analysis. D is the mean change in the corresponding 0–4 criterion score after evidence removal.

Method	Evaluated	D_{cited}	D_{uncited}	D gap	Distractor mass (%)
Post-hoc saliency	352	.27	.19	.08	21.4
Balanced transport	352	.34	.17	.17	16.8
CRaFT (w/o support check)	352	.42	.14	.28	12.1
Full CRaFT	352	.47	.11	.36	7.6

Table 21 Public-data proxy agreement, evidence validity, and constraint-restoration performance with 95% source-film-clustered confidence intervals. Risks are conditioned on issued edits.

Method	Proxy QWK (95% CI)	Evidence valid (95% CI)	Restoration (95% CI)	Eligible/issued	Selective risk (%)
Rubric VLM	.52 [.47, .57]	54.6 [49.2, 60.1]	48.3 [42.8, 54.0]	286/251	23.7
Two-stage	.59 [.54, .64]	66.2 [60.5, 71.6]	60.4 [54.8, 65.9]	286/215	17.1
CRaFT	.66 [.61, .70]	77.1 [72.0, 81.5]	72.0 [66.2, 77.1]	286/181	11.8

Table 22 Calibration and edit-difference coverage under film-clustered evaluation. Nominal proxy-score interval coverage is 90%, and score widths use the 0–4 scale.

Data unit	n_{cal}	n_{test}	Score coverage (%)	Mean width	Edit-difference coverage (%)
Movie scenes (by film)	410	762	90.8	.93	87.1
AVE public sequences (by film)	116	224	91.6	1.18	84.9
MovieNet public scenes (by film)	84	103	90.3	1.25	88.2

4.10 Evidence faithfulness and robustness analyses

Additional tests assess the sensitivity of rubric-aligned evidence, selective reliability, subgroup behavior and neutral-edit invariance. All diagnostics use computational proxies derived from publicly available film annotations and controlled perturbation operators.

Evidence erasure and distracting anchors. For each originally cited panel/transition, remove its evidence token and re-score the associated criterion. Compare the reduction D_{cited} with matched uncited removals of the same modality and temporal span. Insert semantically plausible but unrelated rubric anchors and report spurious transport mass and false citation frequency; use paired source-film or project bootstrap confidence intervals. This direct perturbation test measures faithfulness, beyond merely sharing a transport plan.

Cluster-aware uncertainty and selective risk. Construct confidence intervals by resampling whole source films for public clips; never resample correlated frames, shots or edits as independent observations. Freeze abstention thresholds on calibration data, then evaluate risk at matched issued coverage on untouched film-disjoint test subsets. Report nominal and empirical coverage of the conformal score intervals, mean widths and separately calibrated edit-difference interval coverage, with counts of eligible and issued edits.

Creative diversity and stronger comparators. Stratify public film clips by observable source-level attributes such as dialogue density, action intensity, illumination, and temporal ambiguity; report proxy agreement and false restoration flags. Hand-drawn student artwork and learner experience are outside the scope of the present

Table 23 Performance stratified by observable cinematic conditions. Incorrect restoration is measured among issued computational edit suggestions.

Public clip subset	Sequences	Proxy QWK	Issued edits	Incorrect restoration (%)
Dialogue-heavy scenes	114	.63	83	8.4
Action-heavy scenes	121	.71	97	6.9
Low-light shots	67	.57	42	15.7
Ambiguous shot transitions	59	.64	46	11.2

Table 24 Neutral-edit stability measured as absolute rubric-score differences on the 0–4 scale for neutral and unaffected criteria.

Variant	Pairs	Neutral $\lvert\Delta\rvert$	Unaffected $\lvert\Delta\rvert$
Original ranking loss	276	.24	.19
With neutral loss	276	.11	.08

public-film evaluation. The direct instruction-following multimodal critique baseline and task-specific film-attribute model use the same panels, rubric, output limit and access to intent, with matched evaluation units and no test-label leakage. The subgroup denominators are the 114, 121, 67 and 59 sequences and the 83, 97, 42 and 46 issued edits listed in Table 23. Categories can overlap, so their counts are not combined as independent samples.

Neutral-edit invariance. Compare score changes in ground-truth neutral edit pairs before and after introducing Eq. (22); report the absolute criterion-score difference and stability of untouched criteria on held-out edit pairs, with the original perturbation families unchanged.

4.11 Scope of transfer to educational assessment

The public cinematography benchmarks establish a computational testing framework for observable shot attributes, sequence perturbations, proxy ordinality and evidence localization. They do not provide independent teacher-defined grades, student-created hand-drawn boards, or validated measures of usefulness in an actual classroom. Therefore, none of the current results establish alignment with instructor judgments, improved student revisions, or learning gains. Prospective educational validation would require an independently designed study and, where applicable, institutional ethics review and informed consent.

5 Discussion

5.1 Where the methodological contribution lies

The principal modeling distinction is that the same criterion-conditioned evidence representation determines both an assessment and the allowable direction of revision. The ablations separate the impact of prompt coupling (5.4 points of MovieShots F1),

transport (18.7 points of grounding P@3), and intervention consistency (6.1 points of AVE ordering accuracy). The matched independent, staged, and gradient-detached controls in Table 12 isolate optimization coupling beyond a shared module inventory. These are consistent with the architectural hypothesis that an editable, evidence-linked representation is more informative than independent recognition and post-hoc language generation. The paired clustered evaluation in Table 16 specifies the corresponding statistical and split-level checks.

5.2 Educational validity and style pluralism

Professional-film datasets are observational and style-skewed. An "unusual" cut, orientation reversal, or discontinuous gesture may be productive in experimental cinema, and no optimized population average should automatically override a student intention. This motivates conditioning on g, preserving teacher-supplied rubric weights, and returning reasoned uncertainty rather than a single creativity score. It also changes the interpretation of the proposed edit set: these are options for conversation with an instructor, not mandated corrections. An independent creative-diversity audit should examine whether iterative use reduces the range of student styles or disproportionately penalizes non-mainstream cinematic grammar.

Educational and methodological interpretation. Agreement with computational proxy scores, restoration of synthetic corruptions, actual realized edit quality, independent teacher approval and learning gains are distinct estimands. In the terms of educational feedback research [55], useful commentary additionally requires that learners can interpret and act on it; no pre/post learning test is reported here. From a computational perspective, hybrid multimodal fusion and uncertainty calibration must be validated separately from architectural complexity, as illustrated by cross-domain hybrid-model analyses [58]. The cautionary example of external-validity deficits in unrelated medical-image model assessments [59] is methodological only, not evidence about film students. A future multicourse study should examine creative-intent preservation and feedback uptake across diverse assessment contexts.

5.3 Limitations and threats to validity

Public shot labels validate narrow semantic tasks and controlled transformations, whereas actual teacher-rated storyboards would be required to validate educational critique. None were collected in this study. A score predictor trained on automatically constructed corruption pairs may exploit artifacts of the perturbation procedure. Transport mass is an algorithmic explanation mechanism, not a proof of causal visual attention, and entropic assignments can be unstable if anchors are redundant. Conservative bounds are conditional on error calibration and do not certify how students respond. Hand-drawn panels differ from live-action movie frames; additional domain adaptation may introduce new nuisance correlations. These limitations identify areas for external validation and domain-specific testing.

5.4 Reproducibility, ethics, and data governance

All data used in the study derive from publicly released cinema and video-understanding datasets with existing labels, plus computationally generated proxy labels and controlled perturbations. The research did not recruit students, instructors or other participants; did not collect student assignments, private educational records, new interviews or surveys; and did not conduct classroom interventions. Accordingly, there was no new human-participant research procedure for which consent to participate would apply. The assessment is subject to the authors' institutional ethics policies and the source datasets' licenses. The computational study involved no new human-participant procedures.

The experimental protocol is specified through the implementation configuration, rubric anchors, evaluation partitions, and label mapping in Tables 1, 2, 4, 5, and 18. Film-level identity is the primary unit for split isolation and clustered uncertainty analysis. Computational rubric proxies are distinguished from human ratings throughout. The original datasets are governed by the providers' licenses; copyrighted footage is not redistributed.

6 Conclusion

We presented CRAFT, a criterion-grounded approach to multimodal prompt learning for storyboard assessment that unifies rubric-specific evidence transport, ordinal prediction, and constrained counterfactual revision on an editable scene graph. Its differentiating research claim is structural: an assessment should support an evidence path to a criterion and an actionable, verifiable edit, with abstention when that chain cannot be established. The reported cinematography benchmarks show improved shot-attribute recognition, sequence handling, and rubric-proxy agreement over the listed baselines, while ablations support the contribution of paired prompts, unbalanced transport, and edit-sensitive constraints. All reported experimental targets are derived from public cinematography annotations or constructed proxy labels. No teacher agreement, teacher endorsement or student-learning gains were measured. The supplemental computational proxy analyses examine evidence faithfulness, calibration, dataset heterogeneity, and edit stability. Educational effectiveness and teacher agreement remain distinct questions requiring evaluation with instructor-rated storyboards and prospective learning outcomes.

Data Availability

The publicly documented cinematography datasets referenced in this study are available from their original providers, subject to the applicable licenses, access conditions, and copyright restrictions:

- **MovieShots**: official dataset and annotation access page, https://movienet.github.io/projects/eccv20shot.html.
- **FullShots**: authors' original publication and implementation repository, https://www.nature.com/articles/s41598-023-43281-w and https://github.com/litchiar/ShotClassification. The source publication notes that the underlying film videos

cannot be redistributed owing to copyright restrictions, and that some annotation files are available from the corresponding author upon reasonable request. The repository should not be interpreted as providing unrestricted access to the full film dataset.

- **AVE (Anatomy of Video Editing)**: official annotation and preparation repository, https://github.com/dawitmureja/AVE. The original movie clips are not hosted by the authors; availability of third-party video sources may change.
- **MovieNet**: official project and dataset access information, https://movienet.github.io/; the linked OpenDataLab downloads may require registration and acceptance of the provider's usage terms.

MovieGraphs, cited for related-work context rather than as one of the four reported evaluation benchmarks, has an official annotation repository at https://github.com/katha-ai/MovieGraphs-Dataset-CVPR2018. Dataset access remains subject to the providers' license terms and release-specific annotation availability. The analysis uses published annotations and derived computational rubric proxies, with all evaluation targets and data-use restrictions described in the main text. No newly collected student cohort or independent instructor-rating panel formed part of this study.

Code Availability

The model and evaluation pipeline is described in the Methods and Experiments sections, including training, ontology harmonization, film-level splitting, calibration and perturbation analysis. The source implementation is not publicly deposited at the time of submission. The Methods and Experiments sections provide the model specification and evaluation protocols; publication of the complete code, checkpoint revisions, split manifests, and instance- level predictions would further facilitate independent validation.

Ethics Approval and Consent to Participate

This study consisted solely of computational research on pre-existing, publicly released cinematography datasets (MovieShots, FullShots, AVE and MovieNet) and deterministic or model-generated transformations of their permitted annotations and media. No new human participants were recruited, no student coursework or private educational records were collected, and no educator-rating sessions, classroom trials or other human-participant interventions were conducted. Thus, informed consent to participate was not applicable to the reported research. Because the research involved no newly recruited human participants, ethics approval and participant consent were not applicable to the reported computational study, subject to the requirements of the authors' institutions. The original datasets were accessed under the providers' respective licensing and access conditions.

References

[1] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, 2020. URL https://movienet.github.io/projects/eccv20movienet.html.

[2] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *CVPR*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Vicol_MovieGraphs_Towards_Understanding_CVPR_2018_paper.html.

[3] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *ECCV*, 2020. URL https://movienet.github.io/projects/eccv20shot.html.

[4] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *ECCV*, 2022. URL https://arxiv.org/abs/2207.09812.

[5] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[6] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

[7] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale, 2023. URL https://arxiv.org/abs/2303.15389. arXiv preprint.

[8] Jean-Baptiste Alayrac et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. URL https://arxiv.org/abs/2204.14198.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2305.06500.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2304.08485.

[12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022. URL https://arxiv.org/abs/2109.01134.

[13] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhou_Conditional_Prompt_Learning_for_Vision-Language_Models_CVPR_2022_paper.html.

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. URL https://arxiv.org/abs/2203.12119.

[15] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Khattak_MaPLe_Multi-Modal_Prompt_Learning_CVPR_2023_paper.html.

[16] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Khattak_Self-regulating_Prompts_Foundational_Model_Adaptation_without_Forgetting_ICCV_2023_paper.html.

[17] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yao_Visual-Language_Prompt_Tuning_With_Knowledge-Guided_Context_Optimization_CVPR_2023_paper.html.

[18] Enxin Song et al. Moviechat: From dense token to sparse memory for long video understanding. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Song_MovieChat_From_Dense_Token_to_Sparse_Memory_for_Long_Video_CVPR_2024_paper.html.

[19] Yi Wang, Kunchang Li, Xinhao Li, et al. Internvideo2: Scaling video foundation models for multimodal video understanding, 2024. URL https://arxiv.org/abs/2403.15377. arXiv preprint.

[20] Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, et al. Videoprism: A foundational visual encoder for video understanding. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/zhao24f.html.

[21] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/

Li_VidHalluc_Evaluating_Temporal_Hallucinations_in_Multimodal_Large_Language_Models_for_CVPR_2025_paper.html.

[22] Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jung_On_the_Consistency_of_Video_Large_Language_Models_in_Temporal_CVPR_2025_paper.html.

[23] Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gasevic, and Guanliang Chen. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 2024. URL https://doi.org/10.1016/j.caeai.2024.100299.

[24] Da-Wei Zhang, Melissa Boey, Yan Yu Tan, and Alexis Hoh Sheng Jia. Evaluating large language models for criterion-based grading from agreement to consistency. *npj Science of Learning*, 2024. URL https://www.nature.com/articles/s41539-024-00291-1.

[25] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017. URL https://doi.org/10.1007/s11263-016-0987-1.

[26] Yuzhi Li, Tianfeng Lu, and Feng Tian. A lightweight weak semantic framework for cinematographic shot classification. *Scientific Reports*, 2023. URL https://www.nature.com/articles/s41598-023-43281-w.

[27] Yuzhi Li, Feng Tian, Haojun Xu, and Tianfeng Lu. Toward unified and quantitative cinematic shot attribute analysis. *Electronics*, 2023. URL https://www.mdpi.com/2079-9292/12/19/4174.

[28] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024. URL https://aclanthology.org/2024.emnlp-main.342/.

[29] Chaoyou Fu et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024. URL https://arxiv.org/abs/2405.21075. arXiv preprint.

[30] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754. arXiv preprint.

[31] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for

multi-task long video understanding, 2024. URL https://arxiv.org/abs/2406.04264. arXiv preprint.

[32] Rohit Girdhar et al. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. URL https://arxiv.org/abs/2305.05665.

[33] Alexander Kirillov et al. Segment anything. In *ICCV*, 2023. URL https://arxiv.org/abs/2304.02643.

[34] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2304.07193.

[35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

[36] Bin Xiao et al. Florence-2: Advancing a unified representation for a variety of vision tasks. In *CVPR*, 2024. URL https://arxiv.org/abs/2311.06242.

[37] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lee_Multimodal_Prompting_With_Missing_Modalities_for_Visual_Recognition_CVPR_2023_paper.html.

[38] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *CVPR*, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_Efficient_Multimodal_Fusion_via_Interactive_Prompting_CVPR_2023_paper.html.

[39] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_PromptKD_Unsupervised_Prompt_Distillation_for_Vision-Language_Models_CVPR_2024_paper.html.

[40] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Tian_ArGue_Attribute-Guided_Prompt_Tuning_for_Vision-Language_Models_CVPR_2024_paper.html.

[41] Jinyoung Park, Juyeon Ko, and Hyunwoo J. Kim. Prompt learning via meta-regularization. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Park_Prompt_Learning_via_Meta-Regularization_CVPR_

2024_paper.html.

[42] Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Language_Models_as_Black-Box_Optimizers_for_Vision-Language_Models_CVPR_2024_paper.html.

[43] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. In *WACV*, 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Shen_Multitask_Vision-Language_Prompt_Tuning_WACV_2024_paper.html.

[44] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zanella_On_the_Test-Time_Zero-Shot_Generalization_of_Vision-Language_Models_Do_We_CVPR_2024_paper.html.

[45] Fusheng Hao, Fengxiang He, Fuxiang Wu, Tichao Wang, Chengqun Song, and Jun Cheng. Task-aware clustering for prompting vision-language models. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hao_Task-Aware_Clustering_for_Prompting_Vision-Language_Models_CVPR_2025_paper.html.

[46] Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_DPC_Dual-Prompt_Collaboration_for_Tuning_Vision-Language_Models_CVPR_2025_paper.html.

[47] Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. Revisionllm: Recursive vision-language model for temporal grounding in hour-long videos. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hannan_ReVisionLLM_Recursive_Vision-Language_Model_for_Temporal_Grounding_in_Hour-Long_Videos_CVPR_2025_paper.html.

[48] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Number_it_Temporal_Grounding_Videos_like_Flipping_Manga_CVPR_2025_paper.html.

[49] Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. Rewind: Understanding long videos with instructed learnable memory. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/

Diko_ReWind_Understanding_Long_Videos_with_Instructed_Learnable_Memory_CVPR_2025_paper.html.

[50] Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. Bolt: Boost large vision-language model without training for long-form video understanding. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_BOLT_Boost_Large_Vision-Language_Model_Without_Training_for_Long-form_Video_CVPR_2025_paper.html.

[51] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.679/.

[52] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP System Demonstrations*, 2023. URL https://aclanthology.org/2023.emnlp-demo.49/.

[53] Eman Mudhi AlGhamdi, Yuheng Li, Dragan Gasevic, and Guanliang Chen. Leveraging prompt-based llms for automated scoring and feedback generation in higher education. *Computers and Education*, 2026. URL https://doi.org/10.1016/j.compedu.2025.105511.

[54] Omri Rafaeli, Noam Krohn-Borojovich, Itay Marienberg-Milikowsky, and Dan Vilenchik. Mind the gap: word-embedding and multi-layered literary networks. *Digital Scholarship in the Humanities*, 41(Supplement_1):i213–i229, 2026. doi: 10.1093/llc/fqaf112. URL https://doi.org/10.1093/llc/fqaf112. First published online November 11, 2025.

[55] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007. doi: 10.3102/003465430298487. URL https://doi.org/10.3102/003465430298487.

[56] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303. URL https://doi.org/10.1090/mcom/3303.

[57] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022. URL https://arxiv.org/abs/2107.07511.

[58] Igor Garcia-Atutxa, José Martínez-Más, Andrés Bueno-Crespo, and Francisca Villanueva-Flores. Early-fusion hybrid cnn-transformer models for multiclass ovarian tumor ultrasound classification. *Frontiers in Artificial Intelligence*, 8: 1679310, 2025. doi: 10.3389/frai.2025.1679310. URL https://doi.org/10.3389/frai.2025.1679310.

[59] Igor Garcia-Atutxa, Francisca Villanueva-Flores, Ekaitz Dudagotia Barrio, Javier I. Sanchez-Villamil, José Martínez-Más, and Andrés Bueno-Crespo. Artificial intelligence for ovarian cancer diagnosis via ultrasound: a systematic review and quantitative assessment of model performance. *Frontiers in Artificial Intelligence*, 8:1649746, 2025. doi: 10.3389/frai.2025.1649746.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CRaFTFigureGenerationCodeSupplement.zip