From Visual Cues to Spoken Narration: Rethinking Audio Description

Anonymous ACL submission

Abstract

Audio Description (AD) provides spoken narration of visual events during dialogue gaps, making movies accessible to the visually-impaired audiences. The problem requires determining both *what* (which visual event) and *when* (position for inserting the AD) to narrate, to achieve the best user experience. Prior work has largely reduced the problem to video captioning of pre-segmented video clips, i.e., *what* is largely predefined and *when* is ignored entirely. We propose Cue2Narrate, a two-stage pipeline that jointly predicts *what* and *when* to narrate in longer untrimmed movie clips. A dual-head audio-visual localizer predicts two temporally distinct windows per AD utterance: a visual clue window and a spoken narration window. A LoRA-adapted VLM then generates concise ADs from the predicted visual evidence, trained with a Description Ranking Loss that ranks captions (negative samples) of the same frames lower than the GT AD. To benchmark this new problem statement, we introduce LongLSMDC benchmark with up to 8-min movie clips ($\sim$6.5 min on average). On LongLSMDC, Cue2Narrate outperforms video-only and audio-only localization baselines by 5–12 points in avg. mAP, and AD generation under predicted windows comes within $\sim$2 CIDEr of using GT windows. These results establish the first benchmark for multi-segment AD generation on long-form clips.

1 Introduction

Audio Description (AD) is spoken narration that makes movies accessible to blind and low-vision audiences by describing on-screen content not conveyed through dialogue (Benecke, 2007). A trained human describer watches the film, identifies the visual events that matter to the story (actions, expressions, scene changes, character interactions), and writes short descriptions to be voiced over the soundtrack (Packer et al., 2015; DCMP, 2020; Pavel et al., 2020). Since AD is mixed into the existing audio, each description must ideally be inserted into a dialogue "gap", avoiding overlap with speech (Pavel et al., 2020). The describer therefore makes two distinct decisions each time: *what* visual event to describe, and *when* the description should be voiced. The latter is typically done in a dialogue gap that occurs during or slightly before/after the event itself (Pavel et al., 2020; Branje and Fels, 2012; Fryer, 2016). However, professional AD production is slow and expensive, costing roughly \$15–\$75 per finished min, limiting accessibility to a small fraction of released films (3Play Media, 2022, 2025); automating this process is essential to scale AD coverage.

Most prior works on automatic AD generation significantly reduce the problem. First, they often evaluate in an oracle-window setting: the model is given the temporal interval for which a description should be produced (Rohrbach et al., 2017; Han et al., 2023b, 2024; Fang et al., 2025; Xie et al., 2025). Namely, AD boils down to video captioning over pre-defined short clips. Besides, prior works ignore the distinction between the *what* and *when* intervals. Existing AD datasets provide one type of temporal windows: CMD-AD and TV-AD (Han et al., 2024; Xie et al., 2025) align AD text to AudioVault[1] *spoken* slots, while LSMDC and MAD-Eval (Rohrbach et al., 2017; Pini et al., 2019) provide *visually*-aligned short clips with text. Prior works do not highlight this distinction. To sum up, most evaluate *what to say* given an oracle window without the challenge of long untrimmed video, and without predicting *when to actually speak*. The few works that touch on description timing inherit these limitations: AutoAD-II (Han et al., 2023a) binary-classifies pre-supplied speech gaps, and CA3D (Lee et al.,

[1] AudioVault is a community-maintained non-profit repository of movie ADs. See https://audiovault.net.

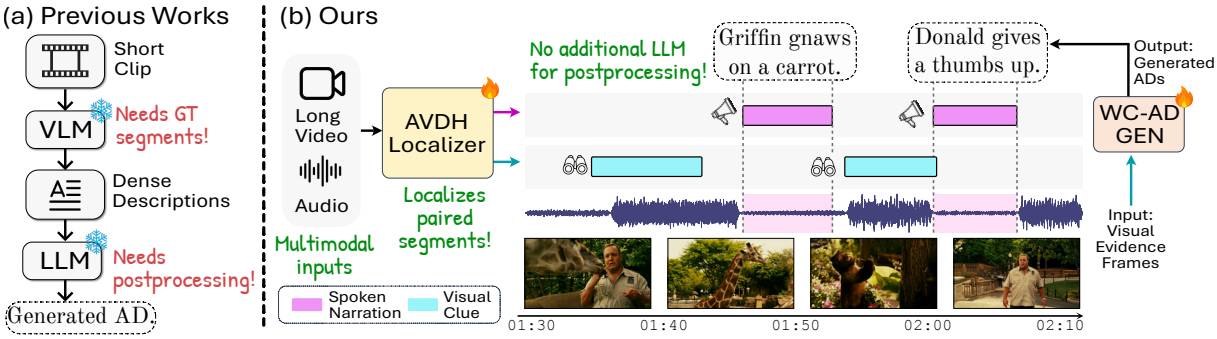


Figure 1: **Cue2Narrate overview.** **(a)** Prior methods caption GT clips with a VLM and then refine the dense descriptions with an LLM into AD-style text. **(b)** Given a long-form clip and the corresponding audio, our AVDH Localizer (Cue2Narrate) predicts paired temporal windows per AD: a visual clue window marking when the event is on screen, and a spoken narration window marking the dialogue gap in which the AD is inserted. Our WC-AD Gen (Cue2Narrate) generates a description from the localized visual window, to be inserted into the predicted spoken window. The figure shows an actual example from LongLSMDC, where these windows are routinely offset.

2025) detects a single event per short clip. Thus, the full AD problem, with visual/spoken cue windows in untrimmed video with multi-segment prediction, has not been tackled so far.

We address these limitations twofold. First, we introduce **LongLSMDC**, a long-form AD benchmark with up to 8-min movie clips (∼6.5 min on average) with dense multi-segment annotations, built on top of (Rohrbach et al., 2017, 2021). Unlike prior benchmarks, LongLSMDC provides both types of windows per AD utterance, matching how human describers produce AD: the visual clue window is paired with the spoken narration window in which it is voiced. Second, we propose **Cue2Narrate** (Fig. 1) , a paired-window AD pipeline for untrimmed video: a dual-head audio-visual localizer (AVDH Localizer) predicts both windows (visual and spoken) per AD utterance, and WC-AD generator produces the description from the predicted localized visual clue window, to be inserted into the spoken narration window.

On LongLSMDC, our localizer outperforms a video-only ActionFormer (Zhang et al., 2022) baseline by 11.8 points in avg. mAP and an audio-only ActionFormer variant by 5.0 points. We also show transfer to CMD-AD and MAD-Eval, where we evaluate on the respective temporal windows. In terms of AD quality, Cue2Narrate improves over its base VLM on both LongLSMDC and CMD-AD; on LongLSMDC the CIDEr under predicted windows is only ∼2 points below the GT-window score. Generated ADs are matched to GT via tIoU on the predicted spoken-narration windows, making Cue2Narrate the first protocol to jointly evaluate *what* and *when* in AD generation.

2 Related Work

Audio Description Generation. Automated AD generation has been studied almost entirely as a text-generation problem given a temporal window. AutoAD (Han et al., 2023b) introduces VLMs for AD generation from a known interval, leveraging text-only AD corpora and visual captioning datasets for pretraining; AutoAD-II (Han et al., 2023a) extends this with character recognition and binary classification of ASR-extracted speech gaps to decide where AD should be inserted, pretraining on AudioVault-AD text and WebVid video-text (Bain et al., 2021); AutoAD-III (Han et al., 2024) releases CMD-AD with AudioVault-aligned spoken slots and introduces HowToAD-style long-video pretraining. Subsequent work focuses on richer generation conditioned on the assumed window, often with external character banks or movie scripts: MM-Narrator (Zhang et al., 2024), NarrAD (Park et al., 2025), UniAD (Wang et al., 2025), AutoAD-Zero (Xie et al., 2024), Shot-by-Shot (Xie et al., 2025), and DistinctAD (Fang et al., 2025). The corresponding benchmarks preserve only one part of the "paired" AD annotation: CMD-AD and TV-AD (Han et al., 2024; Xie et al., 2025) contain *spoken* slots without a visual clue window, while LSMDC and MAD-Eval (Rohrbach et al., 2017; Pini et al., 2019) contain *visual* slots without a spoken narration window. In parallel work (Kala et al., 2025), also operate on few-minute video segments rather than trimmed clips, but focus on QA-based evaluation of *generated* ADs rather than predicting AD timing. None sup-

Table 1: **Audio description datasets** compared across the dimensions needed to predict both AD windows on long-form clips. Prior datasets are either short-clip (LSMDC, MAD-Eval) or contain spoken-only segments on longer clips (CMD-AD, TV-AD; frames available by source). LongLSMDC is the first to support multi-segment prediction of the visual and spoken windows on long-form clips.

Dataset	Source	Domain	#Annotations	Clip length	ADs/clip	Visual	Spoken
LSMDC (Rohrbach et al., 2017)	LSMDC	Movies	108.5K	∼5 sec	1	✓	✗
MAD-Eval (Han et al., 2024)	MAD (LSMDC subset)	Movies	7K	∼5 sec	1	✓	✗
CMD-AD (Han et al., 2024)	CMD + AudioVault	Movies	101K	∼2 min	∼10	✗	✓
TV-AD (Xie et al., 2025)	TV + AudioVault	TV	34K	unknown	-	✗	✓
LongLSMDC (ours)	LSMDC full movies	Movies	83.5K	∼6.5 min	∼40	✓	✓

port joint prediction of the spoken-visual window, and the closest timing-related work, CA3D (Lee et al., 2025), predicts a single window per clip rather than a paired structure. We instead predict the pairs in our Stage 1 on long-form clips by using video and audio modality, supervised by LongLSMDC annotations. Where prior methods generate from an assumed or oversampled visual context, we learn to predict the visual clue window and pair the generated description with the predicted spoken narration window, mirroring how human describers produce AD.

Temporal Localization and Dense Captioning. Temporal action localization (TAL) identifies multi-segment temporal boundaries in untrimmed video with transformer-based architectures (Zhang et al., 2022; Shi et al., 2023) and cross-modal cues (Tian et al., 2018; Senocak et al., 2018, 2023). Dense video captioning extends this to producing descriptions per segment, either as separate (Krishna et al., 2017) or end-to-end (Iashin and Rahtu, 2020; Li et al., 2018; Yang et al., 2023; Islam and Bertasius, 2022) pipelines. Both lines predict visual-only segments anchored to scene changes or action boundaries, with no requirement on the audio track. AD differs structurally: the output is a paired (visual, spoken) annotation, where the spoken window is a dialogue gap that may sit before, during, or after the event being described (Pavel et al., 2020). The visual frame at an AD's spoken window therefore need not match the description's text, placing AD outside the scope of visual-only localization and captioning. Our audio-visual dual-head localizer predicts both windows of the pair, with each head supervised against its own ground-truth window.

3 LongLSMDC Dataset

Our LongLSMDC benchmark is constructed from the full-length movies in LSMDC/LSMDC v2 (Rohrbach et al., 2017, 2021). The commonly used annotations only provide timestamps for visual clue windows. Importantly, we were able to get access to the corresponding spoken narration windows.[2] We segment each movie into non-overlapping clips of up to 8 min (∼6.5 min), assign each AD annotation to the clip containing its temporal midpoint, and convert movie-level timestamps to clip-relative coordinates. The resulting dataset contains 1,995 training clips and 170 validation clips, with an average of ∼40 AD annotations per clip (Table 1. Training annotations span three supervision categories: 37,302 spoken-visual window (y^{vn} and y^{sp}), 21,295 visual-only (y^{vn}), and 20,197 spoken-only (y^{sp}) (∼83k).[3] Within the paired set, 32,715 (∼88%) exhibit a temporal offset between the two windows, confirming that the visual and spoken windows are routinely distinct. The validation set contains 5,006 paired annotations, supporting evaluation against both windows. We additionally create a per-clip character bank, and action-event annotations from each AD sentence via closed-source model (Comanici et al., 2025) to support the Action Score metric (Xie et al., 2025). More details are provided in the Appendix A.

4 Cue2Narrate

4.1 Problem Formulation

Let $X = (\mathcal{V}, \mathcal{A})$ denote a long-form clip with video $\mathcal{V}$ and audio $\mathcal{A}$. The target output is a set of AD annotations

$$\mathcal{Y} = \{(d_i,\, y_i^{vn},\, y_i^{sp})\}_{i=1}^{N}, \qquad (1)$$

where d_i is the AD sentence, $y_i^{vn} = [vn_s^i,\, vn_e^i]$ is the visual clue window interval in which the

[2] Obtained from the original authors.

[3] This is due to various post-processing steps that occured between extraction of the spoken windows and manual alignment to the visual windows carried out in LSMDC.

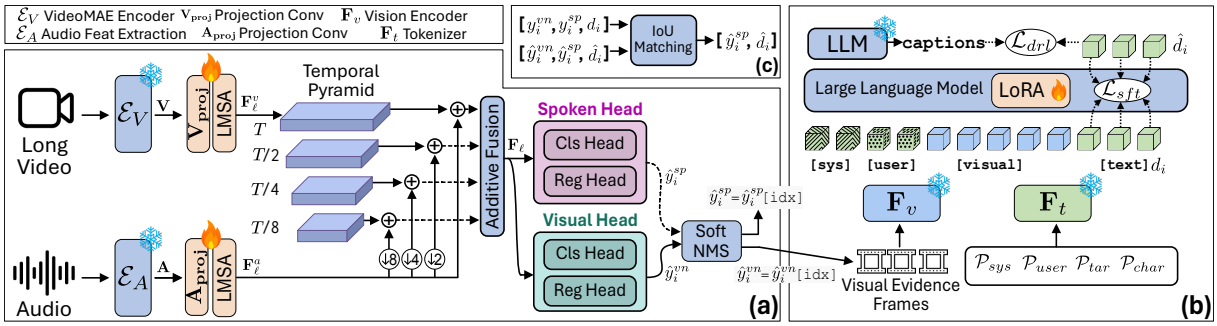


Figure 2: **Cue2Narrate architecture. (a) Localization (Stage 1).** Frozen video ($\mathcal{E}_V$) and audio ($\mathcal{E}_A$) encoders extract features that are projected, refined with self-attention (LMSA), and encoded into a 4-level temporal pyramid ($T, T/2, T/4, T/8$). The streams are combined by additive fusion. Two independent heads the Spoken Head and Visual Head, each with classification and regression branches predict the spoken narration window and visual clue window, respectively. Soft-NMS yields final segments; predicted spoken windows ($\hat{y}_i^{sp}[\texttt{idx}]$) drive evaluation matching, while predicted visual windows ($\hat{y}_i^{vn}[\texttt{idx}]$) condition Stage 2. **(b) Generation (Stage 2).** For each predicted visual clue window, visual evidence frames are encoded into $\mathbf{F}_v$. A structured prompt $\mathcal{P} = [\mathcal{P}_{sys}, \mathcal{P}_{user}, \mathcal{P}_{target}, \mathcal{P}_{char}]$ fed to tokenizer $\mathbf{F}_t$. A frozen LLM with LoRA-adapted projection is trained with $\mathcal{L}_{sft}$ on ground-truth ADs and $\mathcal{L}_{drl}$ against a caption sampled from a frozen LLM on the same frames, generating the AD sentence $\hat{d}_i$. **(c) Evaluation.** Under predicted windows, generated ADs must be matched to ground-truth annotations before scoring; we match on spoken narration window tIoU because that is the side AD references are aligned to in human annotation practice.

described event is visible on screen, and $y_i^{sp} = [sp_s^i,\ sp_e^i]$ is the spoken narration window interval in which d_i is delivered during a dialogue gap. At test time, both the temporal windows and the sentence content are unknown, as well as the number of AD annotations N. Our method factorizes this problem into two stages. Stage 1 predicts paired temporal windows $(\hat{y}_i^{vn}, \hat{y}_i^{sp})$ from the raw clip X. Stage 2 generates a sentence $\hat{d}_i$ from the visual evidence inside the predicted visual clue window, $\hat{y}_i^{vn}$. The final delivered AD is $(\hat{d}_i, \hat{y}_i^{sp})$: the visual clue window determines *what* is described, and the spoken narration window determines *when* that description is inserted.

4.2 Stage 1: Audio-Visual Dual-Head (AVDH) Localizer

Stage 1 localizes AD slots in a long-form clip by predicting, for each instance, a visual clue window window y_i^{vn} and a spoken narration window window y_i^{sp}. In LongLSMDC, 88% of paired annotations exhibit a temporal offset between these two windows; the visible event and the spoken narration differ in onset, duration, or both. We predict them with two independent heads over a shared audio-visual representation, each supervised against its own GT windows. We extract visual features $\mathbf{V} \in \mathbb{R}^{T\times 768}$ with a VideoMAE encoder (Tong et al., 2022) and audio features $\mathbf{A} \in \mathbb{R}^{T\times 130}$ a 128-dimensional log-mel spectrogram resampled to the visual timeline, Fig. 2 (a) The audio stream directly exposes speech activity and silence patterns, critical cues for predicting when a narration slot is available. The visual features are projected and refined with temporal self-attention, then encoded into a multi-scale temporal pyramid $\{\mathbf{F}_\ell^v\}_{\ell=1}^L$ via a ConvTransformer backbone, covering AD events of different durations. The audio features are projected to d_h=512, refined with lightweight temporal self-attention, and average-pooled to each pyramid level, yielding $\{\mathbf{F}_\ell^a\}_{\ell=1}^L$. We fuse the two modalities by pyramid-level addition:

$$\mathbf{F}_\ell = \mathbf{F}_\ell^v + \mathbf{F}_\ell^a. \tag{2}$$

This design is motivated by the complementary strengths of the two modalities (Figs. 3, 4): audio dominates spoken narration window localization where silence patterns are the primary cue, while video dominates visual clue window localization where scene changes and character actions matter Both heads operate on the fused pyramid $\{\mathbf{F}_\ell\}_{\ell=1}^L$ but predict independently. Each head contains a focal classification branch (Lin et al., 2017) for foreground/background scoring and a regression branch that predicts non-negative left/right offsets (δ_l, δ_r) from each temporal anchor t. For pyramid level ℓ with stride s_ℓ, the predicted interval is

$$[t - \delta_l \cdot s_\ell,\ t + \delta_r \cdot s_\ell]. \tag{3}$$

The visual clue window head predicts y_i^{vn}; the

spoken narration window head predicts y_i^{sp}. This allows the model to represent windows with different durations and arbitrary temporal offsets. For spoken classification, we add a learned temporal-position prior over normalized clip time, giving the head a coarse prior over where narration slots tend to occur in a clip. Further, Soft-NMS (Bodla et al., 2017) duplicate predictions by down-weighting overlapping candidates by their tIoU rather than removing them.

Training objective. The two heads are trained jointly:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{cls}}^{vn} + \mathcal{L}_{\text{reg}}^{vn} + \lambda_{sp}\big(\mathcal{L}_{\text{cls}}^{sp} + \mathcal{L}_{\text{reg}}^{sp}\big)\,, \quad (4)$$

with λ_{sp}=1. Both classification terms use sigmoid focal loss (Lin et al., 2017); both regression terms use a 1D center-based DIoU loss. LongLSMDC annotations supervise the heads asymmetrically: visual clue window-only annotations train only the visual head, spoken narration window-only annotations train only the spoken head, and spoken-visual window annotations train both. The paired subset is the key supervision that prevents the two heads from collapsing to a single shared window.

4.3 Stage 2: Window-Conditioned AD Generator (WC-AD GEN)

Stage 2 maps a localized visual clue window window to a single-sentence audio description, Fig. 2 (b). Beyond describing the correct event, the generator must produce narration in the concise AD style rather than captioning the whole clip in great detail. Standard cross-entropy fine-tuning teaches the model *what* to say but not *how*: AD training data is too sparse to overwrite the pretraining prior towards long, descriptive captions. We address this with a Description Ranking Loss (DRL) that suppresses the caption prior during training. We build Stage 2 on Qwen models (Bai et al., 2025) as the backbone, adapting it with LoRA (r=16, α=32, dropout 0.05) on the language model and multimodal projection while keeping the vision encoder frozen. Frames are sampled uniformly from the visual clue window window (F=32 at training, F=64 at inference) and encoded by the frozen vision encoder, then aligned via a trainable projection layer. For each annotation i, the model receives a prompt $\mathcal{P}_i$ containing: (i) a task instruction for one-sentence AD generation, (ii) visual tokens from y_i^{vn}, (iii) optional character-bank identities, and (iv) target-window context. The spoken narration window window is not used as visual evidence; it determines when the generated sentence is inserted.

Loss functions. Let $y_i^+ = d_i$ denote the ground-truth AD and $\mathbf{V}_i$ the sampled frames from y_i^{vn}. The supervised fine-tuning loss is

$$\mathcal{L}_{\text{sft}} = -\sum_t \log P(y_{i,t}^+ \mid y_{i,<t}^+,\, \mathbf{V}_i,\, \mathcal{P}_i). \quad (5)$$

We additionally compute a ranking loss to suppress the caption prior. For each training sample, a negative y_i^- is precomputed by prompting a frozen LLM on $\mathbf{V}_i$ with a simple prompt (no AD framing) and cached offline. We compute mean per-token log-likelihoods for the ground-truth and the negative under the same model and prompt:

$$\begin{aligned} s_i^+ &= \tfrac{1}{|y_i^+|}\sum_t \log P(y_{i,t}^+ \mid y_{i,<t}^+,\, \mathbf{V}_i,\, \mathcal{P}_i), \\ s_i^- &= \tfrac{1}{|y_i^-|}\sum_t \log P(y_{i,t}^- \mid y_{i,<t}^-,\, \mathbf{V}_i,\, \mathcal{P}_i), \end{aligned} \quad (6)$$

where prompt tokens and padding positions are excluded by the label mask. DRL encourages higher normalized likelihood for the concise AD than for the "vanilla" caption:

$$\mathcal{L}_{\text{drl}} = -\log \sigma\Big(\tfrac{s_i^+ - s_i^-}{\tau}\Big)\,. \quad (7)$$

Since the negative is generated by the backbone we fine-tune from, DRL directly targets the residual pretraining prior that cross-entropy alone cannot suppress. Thus, the full Stage 2 objective is:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{sft}} + \lambda\,\mathcal{L}_{\text{drl}}, \qquad \lambda{=}0.2,\ \tau{=}0.5. \quad (8)$$

During training, Stage 2 uses ground-truth visual clue window windows; at inference, predictions from Stage 1 are used instead.

4.4 Joint Inference

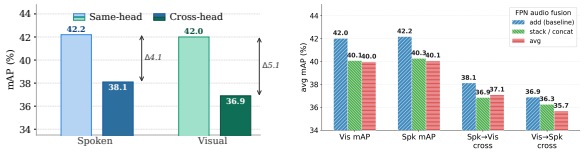
The two stages couple naturally at test time: Stage 1 predicts *where* to look and narrate, Stage 2 generates *what* to say exactly, the same decisions a human describer makes. At test time, Stage 1 predicts paired windows $\{(\hat{y}_i^{vn}, \hat{y}_i^{sp})\}_{i=1}^{\hat{N}}$ from the raw clip. For each pair, Stage 2 generates $\hat{d}_i$ from $\hat{y}_i^{vn}$, conditioned on the previously generated sentence. The final output is:

$$\{(\hat{d}_i,\, \hat{y}_i^{sp})\}_{i=1}^{\hat{N}}, \quad (9)$$

with *no ground-truth boundaries used at any point*. The total cost is one Stage 1 forward pass and one VLM decoding per AD slot.

Table 2: **Stage 1 localization and transfer.** Average mAP over tIoU 0.2–0.6. All methods are trained on LongLSMDC, where Visual and Spoken report each head's mAP against its own GT window, and **Avg.** is their mean. Transfer to CMD-AD reports spoken narration window mAP and to MAD-Eval reports visual clue window mAP. ZS = zero-shot, FT = fine-tuned on the target dataset. "–" indicates not evaluated.

Method	LongLSMDC			CMD-AD		MAD-Eval
	Visual	Spoken	Avg.	FT	ZS	ZS
Spectrogram	1.4	1.3	1.4	1.7	1.7	23.1
ActionFormer	32.0	28.7	30.3	–	–	63.7
Audio-ActionFormer	35.1	39.1	37.1	39.2	25.0	–
Cue2Narrate	**42.0**	**42.2**	**42.1**	**41.0**	**26.5**	**70.0**



(a) (b)

Figure 3: **Stage 1 Dual-Head Analysis.** (a) Each head evaluated against its own GT (**Same-head**) and against the other head's GT (**Cross-head**); the ∼4–5 point drop confirms the visual and spoken heads learn distinct temporal targets rather than a shared window. (b) Audio-fusion strategies (add, stack, avg) compared on visual mAP, spoken mAP, and the two cross-head settings. Element-wise addition consistently wins, justifying our fusion choice.

5 Experiments

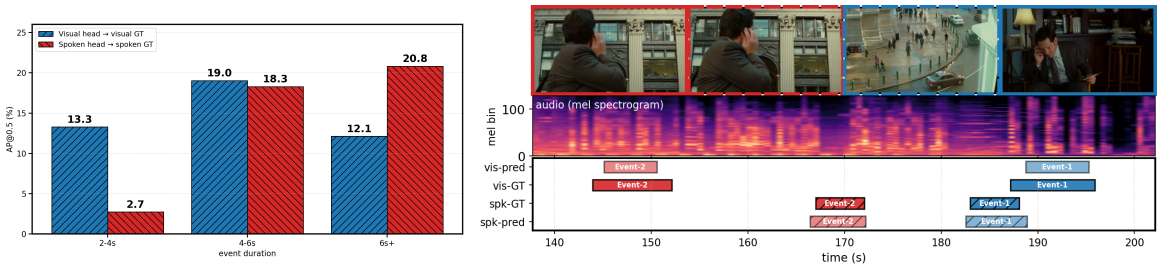
Datasets. For Stage 1, we train on LongLSMDC and evaluate on LongLSMDC, CMD-AD (Han et al., 2024), and MAD-Eval (Han et al., 2024). For Stage 2, we only use the datasets with spoken narration window's annotations. More dataset and metric details in Appendix A.

Joint evaluation. When the AD Generator is conditioned on Stage 1 predictions, Stage 2 generates AD text from predicted visual clue window, while temporal matching for metrics is performed on the paired predicted spoken narration window. Concretely, the predicted visual clue window drives AD generation only, while the predicted spoken narration window drives temporal matching to GT spoken annotations. A prediction is matchable only against GT spoken segments from the same video with tIoU > 0.3; among multiple matches to the same GT segment, greedy one-to-one matching retains the highest tIoU. Text metrics (CIDEr (Vedantam et al., 2015), Action Score (Xie et al., 2025), Recall (Han et al., 2023b), LLM-Eval (Han et al., 2024; Zhang et al., 2024)) are computed on matched pairs only, reducing false positives from unmatched localization errors and keeping evaluation aligned with the spoken-side annotation protocol.

5.1 Localization Results

Table 2 reports localization on LongLSMDC with zero-shot transfer to CMD-AD and MAD-Eval; we also report the fine-tuning results for the former; no additional pretraining data is used. Cue2Narrate achieves 42.0 visual and 42.2 spoken mAP on LongLSMDC, outperforming ActionFormer by 11.8 points and the audio-only variant by 5.0 points. The gain over ActionFormer confirms that audio is a critical modality: silence patterns directly signal narration slot availability, information unavailable to a video-only model. The residual gap between audio-only and the full model reflects the complementary role of video, which identifies which silences correspond to narratively salient events. On CMD-AD, Cue2Narrate reaches 41.0 spoken mAP (FT), above *Audio*-ActionFormer (39.2). On MAD-Eval, it achieves 70.0 visual mAP, above ActionFormer (63.7). CA3D (Lee et al., 2025), the only prior AD localizer, reports 65.3 F1 on MAD-Eval; our model reaches 90.0 F1 on the same split, confirming the gain is not a metric artifact. As a closed-source ablation, we probe Gemini 2.5 Pro (Comanici et al., 2025) for localization via prompted timestamp prediction; despite its strong video understanding, it fails to produce reliable temporal boundaries, highlighting that precise AD localization requires dedicated training (per-tIoU breakdowns in Appendix B).

Dual-head design and fusion strategy. We ablate the two key design choices: The dual-head structure and how the audio stream is fused into the visual FPN. Fig. 3a shows the heads are not redundant: each is strong on its own target but degrades by 4–5 points when evaluated against the other head's GT. The drop reflects the fact that paired visual and spoken windows in LongLSMDC are routinely offset in time (Sec. 3); a single shared head would be forced to predict one window or an average of the two, losing the offset that defines AD timing. The dual-head design preserves it. Fig. 3b compares three strategies for combining the audio stream with the visual FPN: addition (ours), stack/concat, and avg. Addition wins



(a) (b)

Figure 4: **Modality Contribution.** (a) Per-instance AP@0.5 on paired annotations bucketed by event duration, for the visual head against the visual GT window (blue) and the spoken head against the spoken GT window (red). (b) Qualitative example with mel-spectrogram and frame strip with vis-pred/vis-GT and spk-pred/spk-GT tracks, illustrating the temporal offset between paired visual and spoken windows that Cue2Narrate recovers.

consistently across all four settings, with concat and avg trailing by 1.5–2 mAP. Concat forces the model to learn a new joint representation at each pyramid level, diluting the modality-specific signal; avg dampens both streams symmetrically. Addition keeps each modality's contribution intact and lets each head exploit them independently, audio for silence cues, video for visual salience, which is exactly the complementarity the dual-head design was set up to capture.

Per-instance and qualitative evidence: each modality dominates its regime. Audio reliably finds the dialogue gaps where narration can be inserted but cannot say which gaps are worth narrating; video supplies the complementary signal — shot changes, character actions, and on-screen events that distinguish a narratable moment from an arbitrary pause. Fig. 4a shows this split at the instance level on paired annotations bucketed by event duration. The spoken head (red) is weak at 2–4 s (AP@0.5 = 2.7) where dialogue gaps are too short to disambiguate but dominates at 6 s+ (20.8) where the gap structure is unambiguous: Its strength scales with the extent to which the gap is observable. The visual head (blue) is comparatively flat across durations (13.3 vs. 12.1), Since visual evidence accumulates over scene context rather than gap length, a short and a long event can carry a comparable visual signal. The two modalities cover non-overlapping regimes. Fig. 4b makes this concrete on a paired clip with two events: the mel-spectrogram shows a dialogue dip in each, which the spoken head (`spk-pred`) aligns with, while the visual head (`vis-pred`) locks onto the on-screen action visible in the frame strip. The visual clue window and spoken narration window for the same event are temporally offset; the action and its narration do not coincide, and Cue2Narrate recovers both sides of the offset.

Table 3: **Stage 2 generation on LongLSMDC and CMD-AD.** Results use ground-truth spoken narration window segments; LongLSMDC is evaluated on paired data with spoken narration window windows, and CMD-AD uses CharBank when available. Code is unavailable for AutoAD-III and DistinctAD, so only their reported CMD-AD numbers are included. TF = training-free; C=CIDer; AS=Action Score; R@1/5=Recall@1/5.

Method	TF	LongLSMDC			CMD-AD (Zero-shot)		
		C	AS	R@1/5	C	AS	R@1/5
AutoAD-III	✗	–	–	–	25.0	31.5	31.2
DistinctAD	✗	–	–	–	22.7	–	33.0
Uni-AD	✗	–	–	–	21.8	–	–
AutoAD-Zero	✓	14.2	–	31.3	17.7	25.5	26.9
Shot-by-Shot	✓	16.9	25.3	28.4	**26.3**	28.4	33.0
Qwen2.5-VL-7B	✗	27.4	42.6	46.2	18.9	35.0	33.2
+ **Cue2Narrate**	✗	28.5	41.9	44.5	20.5	35.8	32.6
Qwen3-VL-8B	✗	34.6	45.2	49.2	21.6	36.1	34.3
+ **Cue2Narrate**	✗	**36.2**	**46.3**	**49.6**	22.7	**37.2**	**34.8**

5.2 Generation Results

We further extend to the *what* component of AD, generating the sentence that describes the event.

Generation under predicted and GT spoken windows. Tables 4 and 3 report generation quality in two settings. We evaluate on paired spoken annotations. **S1→S2** (Table 4) Cue2Narrate generates from Cue2Narrate predicted visual windows and is scored against GT ADs matched through the predicted spoken window (Sec. 5). **GT→S2** (Table 3) replaces predicted windows with GT spoken segments, removing Cue2Narrate localization error from the evaluation and measuring generator quality in isolation. In both settings, Cue2Narrate improves over the base model on both backbones and both datasets: on LongLSMDC, +1.5 CIDEr for Qwen3-VL (32.5 → 34.0) under predicted win-

Table 4: **Stage 1+2 generation on LongLSMDC and CMD-AD.** Results are computed under Stage 1 predicted segments using tIoU@0.3 matching. Results on paired data. TF = training-free; C=CIDer; AS=Action Score; R@1/5=Recall@1/5.

Method	TF	LongLSMDC			CMD-AD (Zero-Shot)		
		C	AC	R@1/5	C	AC	R@1/5
AutoAD-Zero	✓	10.9	33.1	27.4	10.1	**30.9**	25.3
Qwen2.5-VL-7B	✗	27.5	41.9	45.5	7.5	23.4	24.8
+ **Cue2Narrate**	✗	28.7	41.2	43.4	9.2	25.1	24.9
Qwen3-VL-8B	✗	32.5	45.0	**48.9**	8.8	24.8	25.2
+ **Cue2Narrate**	✗	**34.0**	**45.8**	48.5	**11.5**	27.8	**26.0**

Method	C	LLM
Gemini-2.5 Pro	17.3	2.80
Qwen2.5 + Cue2Narrate	28.5	2.43
Qwen3 + Cue2Narrate	**36.2**	2.68

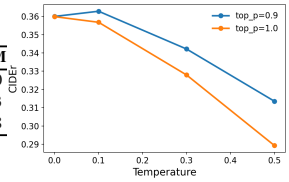


Figure 5: **Stage 2: LongLSMDC spoken windows. Left:** comparison against Gemini-2.5 Pro. **Right:** decoding-strategy ablation for Cue2Narrate greedy decoding performs best among finished.

dows and +1.6 CIDEr (34.6 → 36.2) under GT windows. On CMD-AD, the gain transfers zero-shot (8.8 → 11.5 predicted; 21.6 → 22.7 GT). *The narrow 2-point gap between GT→S2 and S1→S2 on LongLSMDC*: it shows that Cue2Narrate localizes well enough that the end-to-end pipeline operates near the generator's oracle-window ceiling. On CMD-AD, the generator recovers when oracle input segments are provided (22.7 CIDEr) but drops sharply under predicted windows (11.5 CIDEr). Stage 1 still predicts a visual window for CMD-AD, but CMD-AD provides no GT visual annotations, so the generator was never exposed to CMD-AD distribution of visual windows; the predicted visual windows come from a localizer trained on LongLSMDC, while CMD-AD clips are defined by narration-slot timing rather than visual events. CMD-AD therefore serves as a transfer benchmark.

Qualitative analysis on dense, long-form clips. Fig. 6 shows a CMD-AD clip with two AD events: Cue2Narrate recovers both spoken windows and generates contextually relevant descriptions, though phrasing diverges from the reference (e.g., *"Ace stands and watches Chris walk away"* vs. GT *"Gordy pulls the hammer"*) a common pattern when multiple plausible visual subjects share the frame. Dense, multi-event examples from LongLSMDC are in Appendix G.

GT 1: Gordy pulls the hammer.

Pred 1: Ace stands and watches Chris walk away in the woods.

GT 2: Ace swallows and glances at Charlie.

Pred 3: Ace lowers his gaze and shifts his weight.

Figure 6: **Qualitative example on CMD-AD.** Top: sampled frames. Bottom: Pred segments (blue) vs. GT (black). GT/Pred AD sentences are listed below.

Comparison against closed-source models and decoding ablation. Fig. 5 probes two questions about Cue2Narrate. **Left:** how does Cue2Narrate compare to a strong closed-source VLM? Gemini-2.5 Pro (Comanici et al., 2025), evaluated in the **GT→S2** setting, trails both Cue2Narrate backbones on CIDEr but matches them on LLM-Eval; it produces coherent text but does not adopt the AD style that Cue2Narrate's Description Ranking Loss explicitly trains for , confirming that AD generation rewards task-specific adaptation. Tables 4 and 3 isolate this effect: the `+Cue2Narrate` rows, which add $\mathcal{L}_{drl}$ on top of SFT, consistently improve over their SFT-only counterparts on CIDEr, showing that DRL provides gains beyond standard fine-tuning. **Right:** how sensitive is Cue2Narrate to decoding choices? Greedy decoding wins among the settings tested, indicating Cue2Narrate's output distribution is sharp enough that sampling rarely finds better phrasings.

6 Conclusion

We re-framed automatic Audio Description as the joint prediction of *what* to narrate and *when* to narrate it, and introduced LongLSMDC to benchmark this problem on long-form movie clips. Cue2Narrate pairs a dual-head audio-visual localizer with a window-conditioned generator. The dual head separately predicts where the event is visible and where its narration is voiced, recovering the offset between the two that defines AD timing. The window-conditioned generator then produces the AD sentence from the predicted visual window, to be inserted into the predicted spoken window. We hope LongLSMDC and Cue2Narrate open up multi-segment AD as a tractable problem.

7 Limitations

This work is a first step toward automatic audio description on long-form, untrimmed video, and opens up new directions. Our two-stage pipeline keeps localization and generation separate; jointly training the two stages would let generation feedback shape window prediction and is a natural next step. Stage 2 currently conditions only on the visual clue window — extending the generator to also reason over the spoken narration window (e.g., surrounding dialogue, adjacent ADs) could further improve narrative coherence across consecutive descriptions.

8 Ethical Considerations

Cue2Narrate is intended as an assistive tool to expand AD coverage for blind and low-vision audiences, complementing rather than replacing human describers. Automatically generated ADs can be contextually plausible but can be factually wrong, which is particularly consequential for users who depend on them as their primary access to visual content; deployment should disclose machine generation and support human review. Both the training data and the underlying VLM carry demographic biases that may surface in generated descriptions, and auditing for descriptive fairness is an important direction we leave to future work. LongLSMDC annotations follow LSMDC's research-only terms; the underlying video content is not redistributed.

References

3Play Media. 2022. How much does audio description cost? Company blog post. Lists typical AD production pricing ranges ($15–$75 per minute).

3Play Media. 2025. How to select an audio description vendor. https://www.3playmedia.com/blog/select-audio-description-vendor/. Accessed: 2026-05-25.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*.

Bernd Benecke. 2007. Audio description: phenomena of information sequencing. In *EU-High-Level Scientific Conference Series MuTra 2007 LSP Translation Scenarios: Conference Proceedings*.

Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-nms–improving object detection with one line of code. In *ICCV*.

Carmen J Branje and Deborah I Fels. 2012. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DCMP. 2020. Description Key. https://dcmp.org/learn/captioningkey/624. Accessed: 2026-05-22.

Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B. Chan. 2025. DistinctAD: Distinctive audio description generation in contexts. In *CVPR*.

Louise Fryer. 2016. *An introduction to audio description: A practical guide*. Routledge.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023a. Autoad ii: The sequel-who, when, and what in movie audio description. In *ICCV*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023b. Autoad: Movie description in context. In *CVPR*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. Autoad iii: The prequel-back to the pixels. In *CVPR*.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *CVPR Workshops*.

Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. In *ECCV*.

Divy Kala, Eshika Khandelwal, and Makarand Tapaswi. 2025. What you see is what you ask: Evaluating audio descriptions. In *EMNLP*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *ICCV*.

Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. 2025. Now you see me: Context-aware automatic audio description. In *WACV*.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *CVPR*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.

Jaclyn Packer, Katie Vizenor, and Joshua A Miele. 2015. An overview of video description: history, benefits, and guidelines. *Journal of Visual Impairment & Blindness*.

Jaehyeong Park, Junchel Ye, Seungkook Lee, Hyun W Ka, and Dongsu Han. 2025. Narrad: Automatic generation of audio descriptions for movies with rich narrative context. In *WACV*.

Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*.

Stefano Pini, Marcella Cornia, Federico Bolelli, Lorenzo Baraldi, and Rita Cucchiara. 2019. M-vad names: a dataset for video captioning with naming. *Multimedia Tools and Applications*.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *IJCV*.

Anna Rohrbach and 1 others. 2021. https://sites.google.com/site/describingmovies/download.

Arda Senocak, Junsik Kim, Tae-Hyun Oh, Dingzeyu Li, and In So Kweon. 2023. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In *WACV*.

Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *CVPR*.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. 2023. Tridet: Temporal action detection with relative boundary modeling. In *CVPR*.

Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. 2023. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *CoRR*.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Hanlin Wang, Zhan Tong, Kecheng Zheng, Yujun Shen, and Limin Wang. 2025. Contextual AD narration with interleaved multimodal sequence. In *CVPR*.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Eshika Khandelwal, Gül Varol, Weidi Xie, and Andrew Zisserman. 2025. Shot-by-shot: Film-grammar-aware training-free audio description generation. In *ICCV*.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD-Zero: A training-free framework for zero-shot audio description. In *ACCV*.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *CVPR*.

Chenlin Zhang, Jianxin Wu, and Yin Li. 2022. Actionformer: Localizing moments of actions with transformers. In *ECCV*.

In this supplementary material, we provide a comprehensive analysis supporting our main paper claims through: (1) Dataset and Character-bank Details (Sec. A), (2) fine-grained IOU evaluation for LongLSMDC (Sec. B), (3) Human evaluation (Sec. C), (4) Fine-tuning prompt details for Cue2Narrate (Stage 2) (Sec. F), (5) Examples of negative caption examples (Sec. D), (6) Gemini prompt details for Action-Score metric (Sec. E), (7) additional qualitative results (Sec. G), (8) License For Artifacts and its uses (Sec. H) (9) Model size and budget (Sec. I).

We plan to release LongLSMDC annotations and features to the research community.

A Additional LongLSMDC and Character-bank details

The original LSMDC (Rohrbach et al., 2017) consists of full-length movies (90–180 minutes) with sparsely distributed AD annotations. Processing full-length movies end-to-end is computationally prohibitive for both training and evaluation: a single 90-minute film contains thousands of frames and hundreds of AD slots, making batched temporal localization infeasible at movie scale. We therefore segment each movie into variable-length clips (5–8 minutes), each containing multiple AD annotations, enabling scalable training and dense multi-segment evaluation. We will release LongLSMDC clips, annotations, and character banks to the research community.

Random clip segmentation. For each movie of duration D, we divide it into non-overlapping clips with randomized durations $\delta \sim \mathcal{U}(300\text{s}, 480\text{s})$. Starting from $t = 0$, we create clip $[t, \min(t + \delta, D))$, then advance t to the clip end, continuing until the entire movie is segmented. Random lengths prevent duration-specific biases and create variable clip complexity (1–147 annotations per clip). The few clips shorter than 300 s (<0.2%) are tail segments produced when a movie ends before the sampled duration expires.

Annotation assignment via midpoint rule. Each AD annotation $[s, e]$ is assigned to clip $c = [t_s, t_e)$ if its temporal midpoint $m = (s + e)/2$ satisfies $t_s \leq m < t_e$. This ensures each annotation belongs to exactly one clip and is temporally centered for maximum context. We compute clip-relative coordinates $s' = s - t_s$ and $e' = e - t_s$ for standardized temporal ranges. Annotations on clip boundaries (<1%) are excluded.

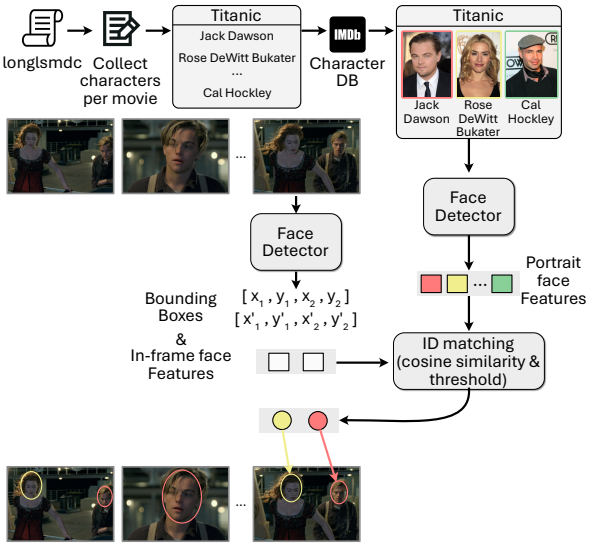


Figure 7: **LongLSMDC character bank construction pipeline.** For each movie, we collect a cast list of character names from an external database. A face detector is applied to movie frames to extract bounding boxes and in-frame face features, which are compared against portrait-level face features for each character via cosine similarity with a fixed threshold, assigning detections to named characters. The resulting per-movie character bank maps frame-level detections to character identities (e.g., Jack Dawson, Rose DeWitt Bukater, and Cal Hockley in *Titanic*), enabling CharBank-conditioned AD generation at inference time.

This transforms ∼125 movies into 2,100+ clips (1,995 training, 170 validation) with over 78,000 training annotations across all annotation categories. Individual clips contain between 1 and 147 AD annotations (mean: 29–40 ADs per clip depending on annotation category), providing diverse density conditions for localization evaluation. Dataset density statistics comparing LongLSMDC and CMD-AD are shown in Figure 9: CMD-AD contains short clips (2–3 min) with 1–5 ADs per clip, while LongLSMDC contains longer clips (up to 8 min) with 29–40 ADs per clip, providing substantially more AD events per movie and better supporting evaluation of long-form, multi-segment AD generation.

Character bank construction. Following AutoAD-Zero (Xie et al., 2024), we construct a per-clip character bank for LongLSMDC to support CharBank-conditioned generation experiments (Figure 7). For each movie, we collect a cast list of character names from IMDb. A face

Figure 8: **Vocabulary comparison between LongLSMDC (Top) and CMD-AD (Bottom) datasets.** LongLSMDC exhibits richer action verb diversity (e.g., "walks", "turns", "looks", "stands") and more varied character descriptors, reflecting its longer clips and broader movie coverage across diverse genres and cinematic styles.

detector is applied to movie frames to extract bounding boxes and in-frame face features, which are then compared against portrait-level face features for each character via cosine similarity with a fixed threshold, assigning frame-level detections to named characters. The resulting per-movie character bank maps detected faces to character identities (e.g., Jack Dawson, Rose DeWitt Bukater, and Cal Hockley in *Titanic*), enabling CharBank-conditioned AD generation at inference time. Unlike AutoAD-Zero, which constructs character banks over ground-truth segments, our banks are built over full movie frames and are therefore compatible with predicted windows. Character metadata is provided per clip for both training and validation splits and will be released alongside the dataset.

Vocabulary analysis. Figure 8 presents word cloud visualizations comparing LongLSMDC and CMD-AD vocabularies. LongLSMDC exhibits richer action verb diversity (e.g., "walks", "turns", "looks", "stands") and more varied character descriptors, reflecting its longer clips and broader movie coverage across diverse genres and cinematic styles. This vocabulary richness supports more comprehensive evaluation of generation models' ability to produce varied, natural language descriptions across different narrative contexts.

Metrics. For *localization* we report mean Average Precision (mAP) at temporal IoU thresholds, averaged over $\{0.2, 0.3, 0.4, 0.5, 0.6\}$. Because LongLSMDC supplies two ground-truth views per opportunity, we report mAP for each head against each view, giving a 2×2 matrix that reveals whether the visual and spoken heads predict the same window or genuinely distinct targets

We follow standard Temporal Action Localization (TAL) evaluation protocols (Tang et al., 2023; Zhang et al., 2022), reporting mean Average Precision (mAP) at temporal IoU thresholds $tIoU \in \{0.2, 0.3, 0.4, 0.5, 0.6\}$. These thresholds are widely used to evaluate TAL methods and assess localization quality at varying precision levels. The temporal IoU (tIoU) is defined as the intersection over union between two temporal windows, equivalent to the 1D Jaccard index.

A predicted segment (t_s^{pred}, t_e^{pred}) matches ground-truth (t_s^{gt}, t_e^{gt}) if their temporal IoU exceeds the threshold:

$$\text{IoU} = \frac{|[t_s^{pred}, t_e^{pred}] \cap [t_s^{gt}, t_e^{gt}]|}{|[t_s^{pred}, t_e^{pred}] \cup [t_s^{gt}, t_e^{gt}]|} \geq \text{threshold} \tag{10}$$

For each ground-truth segment, we follow the greedy matching approach: we find the highest-IoU predicted segment that exceeds the threshold. Matched predictions are considered true positives (TP), unmatched predictions are false positives (FP), and unmatched ground-truth segments are false negatives (FN). Average Precision (AP) is computed per video using the precision-recall curve, and mAP is the mean across all test videos. Given a tIoU threshold, mAP computes the mean of average precision across all videos. We also report average mAP (avg. mAP) by averaging mAP values across all tIoU thresholds $\{0.2, 0.3, 0.4, 0.5, 0.6\}$, providing a single metric that captures performance across different localization precision requirements.

For *generation* we use four complementary metrics: **CIDEr** (Vedantam et al., 2015) for n-gram consensus with reference captions, **Recall@k** ($k = 1, 5$) for embedding-space similarity via retrieval, **Action Score** (Xie et al., 2025) for verb correctness (the most informative AD attribute for the listener), and **LLM-Eval** (LLaMA3-8B (Grattafiori et al., 2024) as judge,

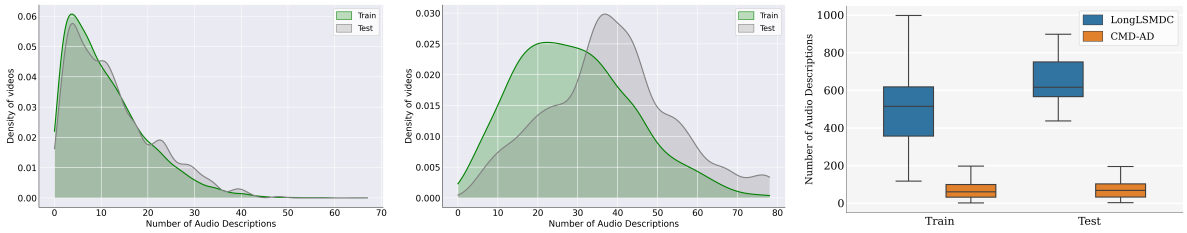


Figure 9: Dataset statistics for CMD-AD and LongLSMDC. (a) CMD-AD contains short clips (2–3 min) with 1–5 audio descriptions per clip. (b) LongLSMDC contains longer clips (up to 8 min) with a much higher density of AD segments (29–40 per clip). (c) Boxplots of the full movie-level distributions show that LongLSMDC provides substantially more AD events per movie than CMD-AD, supporting evaluation of long-form, multi-segment AD generation.

1–5 score, mean reported). All generation evaluations are without CharBank: although character grounding improves AD quality (Han et al., 2023a), per-clip CharBanks (including the one we construct for LongLSMDC, supp.) cannot be reliably supplied under predicted windows.

B Fine-Grained IoU Analysis for LongLSMDC

Figs. 10 and 11 break down mAP by tIoU threshold $\{0.2, 0.3, 0.4, 0.5, 0.6\}$. Three patterns emerge. First, Gemini 2.5 Pro (Comanici et al., 2025) is competitive at loose tIoU on LongLSMDC ($\sim$ 12 mAP at 0.2) but collapses sharply as the threshold tightens, falling below 3 mAP at tIoU 0.5 and effectively zero at 0.6. This indicates that prompted timestamp prediction from a general-purpose VLM can identify approximately when an event occurs but cannot place precise temporal boundaries — exactly the property AD localization requires. Second, on LongLSMDC, Cue2Narrate's margin over the unimodal baselines is modest at tIoU 0.2 but widens at stricter thresholds: at tIoU 0.5, Cue2Narrate reaches 25.3/28.6 vs. video-only's 22.8/17.2 and audio-only's 21.8/25.3. This is consistent with the modality roles observed elsewhere in our analysis: audio carries the dominant cue for spoken-slot timing, while video provides the scene-change and action cues needed to disambiguate which silence corresponds to a narratable event (Table 2). Third, the same pattern holds zero-shot on CMD-AD and MAD-Eval, where Cue2Narrate keeps its lead across all thresholds. On CMD-AD, target-dataset fine-tuning has little effect at loose tIoU ($\sim$ 56 vs. $\sim$ 57 at 0.2) but produces a large gain at strict tIoU (FT 42.7 vs. ZS 12.8 at 0.4), showing that fine-tuning primarily improves boundary precision rather than event recall. The dual-head structure transfers asymmetrically by design: the spoken narration window head supervises CMD-AD-style spoken slots, the visual clue window head supervises MAD-Eval-style visual clips, and each head transfers to the benchmark whose window type it was trained against.

C Human Evaluation

To complement automatic metrics, we run a pairwise preference study comparing Cue2Narrate against Shot-by-Shot (Xie et al., 2025) on 200 randomly sampled LongLSMDC clips. For each clip, 20 evaluators are shown the video alongside two anonymised AD tracks (ours and the baseline, order randomised) and asked to choose the preferred track on three axes: (1) **Coverage of key moments** whether the ADs cover the salient on-screen actions and interactions; (2) **Temporal order & coherence** whether the ADs appear at appropriate moments and follow the story flow; and (3) **Narrative quality & clarity** whether the ADs read as coherent, listener-friendly narration. Fig. 12 reports the percentage of pairwise judgments preferring each system. Cue2Narrate is preferred on every axis, with the largest gap on narrative quality, consistent with the AD-style framing the Description Ranking Loss is designed to encourage.

D Examples of negative caption examples.

To construct the negatives used by DRL (Sec. 4.3), we prompt the frozen Qwen3-VL backbone with a simple prompt (shown in the box below) on each training clip and cache the resulting caption offline. The same prompt format is used across all clips, so the negatives reflect the back-

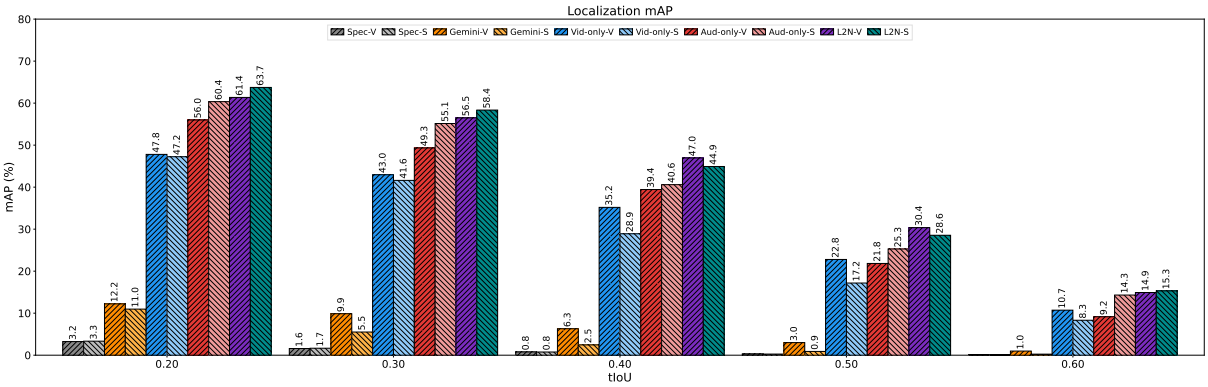


Figure 10: **Per-tIoU localization on LongLSMDC.** Grouped bar plot of mAP at tIoU thresholds $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ (x-axis), with bar height encoding mAP (%). For each threshold, ten bars compare five methods evaluated against both the visual clue window head (**-V**, hatched) and the spoken narration window head (**-S**, solid): Spectrogram (Spec), Gemini 2.5 Pro, video-only ActionFormer (Vid-only), audio-only ActionFormer (Aud-only), and Cue2Narrate (C2N). Gemini is competitive at loose tIoU but collapses at stricter thresholds; Cue2Narrate leads across all thresholds, with the margin over unimodal baselines widening as tIoU tightens.

bone's default captioning prior long, descriptive, scene-summarising rather than any AD-style framing. Two representative pairs of (caption negative, ground-truth AD) are shown in Fig. 13; the detailed-vs-concise gap visible in both examples is what DRL is trained to exploit.

System prompt. You are an intelligent chatbot designed for generating audio descriptions for movies. Audio Description (AD) provides concise spoken narration of important visible events, actions, characters, and scene changes in a movie. The prompt includes several short examples of valid AD sentences.

User prompt. `<video>` Watch the following movie clip and generate an Audio Description for it.

Constraints. The generated AD should be exactly one sentence, use present tense, describe only visible information, avoid hallucination, focus on important people, actions, and interactions, and avoid naming characters unless their identity is visually explicit or provided.

E Gemini prompt details for Action-Score metric

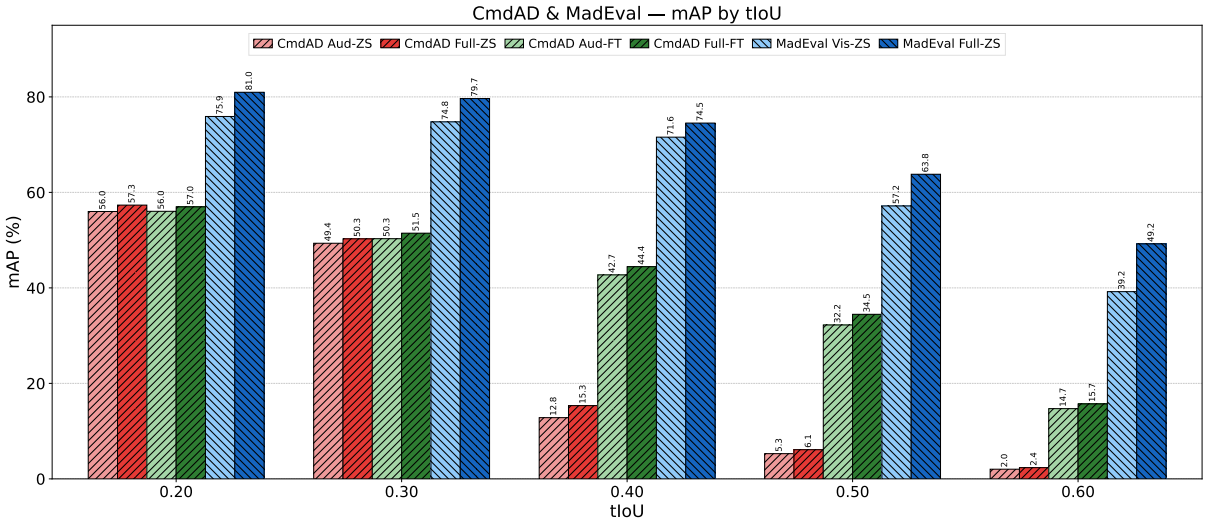


Figure 11: **Per-tIoU transfer to CMD-AD and MAD-Eval.** Grouped bar plot of mAP at tIoU thresholds $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ on the two transfer benchmarks. CMD-AD reports spoken narration window-window mAP with audio-only (**Aud**, red) and the full Cue2Narrate (**Full**, dark red) under both zero-shot (**ZS**, hatched) and fine-tuned (**FT**, solid) settings. MAD-Eval reports visual clue window-window mAP for video-only (**Vis-ZS**, light blue, hatched) and the full Cue2Narrate (**Full-ZS**, dark blue, solid), evaluated zero-shot. On CMD-AD, ZS and FT are nearly identical at loose tIoU but FT pulls sharply ahead at stricter thresholds, showing that target-dataset fine-tuning primarily sharpens boundary precision. Cue2Narrate's lead persists across all thresholds and both transfer benchmarks.

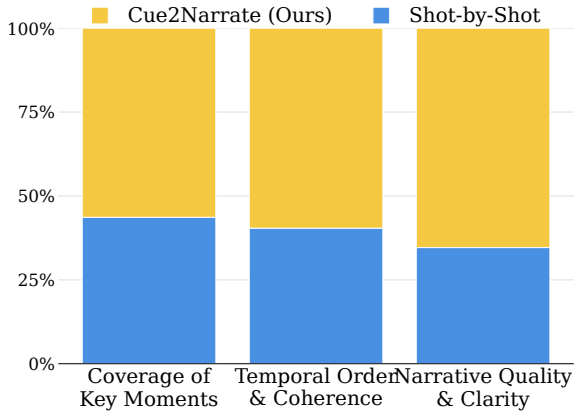


Figure 12: **Human evaluation on LongLSMDC.** Pairwise preference between Cue2Narrate (yellow) and Shot-by-Shot (Xie et al., 2025) (blue) across three dimensions. Evaluators preferred Cue2Narrate on all three axes, with the largest margin on Narrative Quality & Clarity (∼65% vs. 35%).

Example 1 `train_3001_21_JUMP_STREET_clip0_176.321_178.664.mp4`

Caption negative: *The video begins with a close-up shot of a young man with short, light-colored hair, wearing a white shirt. He is sitting on a bench, looking off to the side with a serious expression. The background shows a brick wall and a green bench, suggesting an outdoor setting, possibly a school or public area. The scene then transitions to a wide shot of the exterior of a building, which appears to be a school or institutional facility, with a glass door marked "School of Business." Two individuals are walking away from the entrance.*

Ground-truth AD: Sitting on a nearby bench, the nerd looks over.

Example 2 `train_3001_21_JUMP_STREET_clip0_179.061_182.119.mp4`

Caption negative: *The video opens with a wide shot of the exterior of a school building, where two students are standing outside. One student, wearing a purple and white jacket, is leaning against the wall with their head down, appearing to be in distress or deep thought. The camera then zooms in on the student, who continues to lean against the wall, occasionally wiping their face with their hand, suggesting they might be crying.*

Ground-truth AD: The jock slumps down on a bench on the other side of two columns.

Figure 13: **Caption negatives used by the Description Ranking Loss.** For each training clip, a frozen LLM Qwen3-VL produces a long, descriptive caption (italicized) that DRL pushes the model away from in favor of the concise ground-truth AD.

System Prompt

<role>

You are an intelligent chatbot designed for removing character information of a sentence.

</role>
<instruction>

Here 's how you can accomplish the task : You should replace all character information including names , roles , and jobs into pronouns (e.g., he , she , they , her , him , them). Note , objects, locations , and animals are not counted as character information and should be kept as - is .

</instruction>
<output_format>

You should output the result in JSON format WITHOUT providing ANY more sentences at the beginning or at the end .

</output_format>

User Prompt

Please read the sentence below that describes a video clip:
Input sentence: <text_gt>
Replace all character information including names , roles , and jobs into pronouns (e.g., he , she , they , her , him , them). Note , objects, locations , and animals are not counted as character information and should be kept as - is .
Here are some examples:

- Example1: Input sentence : Spicoli watches Mr . Hand pass out the schedule. Ouput : He watches him pass out the schedule.
- Example 2: Input sentence : Waiting for a reply , the inspector has a look of smug satisfaction as he combs his neat moustache. Output: Waiting for a reply , he has a look of smug satisfaction as he combs his neat moustache.
- Example 3: Input sentence : Emmerich 's eyebrows twitch as he watches her. Output : His eyebrows twitch as he watches her.
- Example 4: Input sentence : Inside is a second pair of doors. Output : Inside is a second pair of doors.
- Example 5: Input sentence : The blonde saunters over to him in her striped pantsuit and leans in for a kiss. Output : She saunters over to him in her striped pantsuit and leans in for a kiss.

System Prompt

<role>

You are an intelligent chatbot designed for decompose the sentence into subsentences .

</role>
<instruction>

Here 's how you can accomplish the task : You should split (rewrite if needed) the sentence into subsentences , each containing only one action phrase (i.e., verb phrase).

</instruction>
<output_format>

You should output your answer in JSON format WITHOUT providing ANY more sentences at the beginning or at the end .

</output_format>

Then the sentence with the removed character is used as input, and the ground truth actions are extracted:

User Prompt

Please read the sentence below that describes a video clip:
Input sentence : <text_wo_char>
Split and rewrite the sentence into subsentences , each containing only one action (i.e.,verb phrase) and preserving all other information (e.g., locations , time , affections , etc .) . Do not output repeating actions.
Examples :

- Example1: Input sentence : He watches him pass out the schedule . Subsentences : [He watches him ., He passes out the schedule .]"
- Example2: Input sentence: Waiting for a reply , he has a look of smug satisfaction as he combs his neat moustache . Subsentences : [He waits for a reply ., He has a look of smug satisfaction ., He combs his neat moustache .]
- Example3: Input sentence : He swings in front of Kingpin , then bounces off a building and kicks the criminal into the air . Subsentences : [He swings in front of him ., He bounces off a building ., He kicks him into the air .]
- Example4: Input sentence : His eyebrows twitch as he watches her . Subsentences : [His eyebrows twitch ., He watches her .]
- Example5: Input sentence : Inside is a second pair of doors . Subsentences : [Inside is a second pair of doors .]

F Cue2Narrate (Stage 2) Fine-tuning Prompts

For fine-tuning the Cue2Narrate (Stage 2) module, we employ structured prompts to generate audio descriptions conditioned on localized temporal boundaries. Prompts `System Prompt` and `User Prompt` show the complete prompt design used during fine-tuning.

System Prompt

You are an intelligent chatbot designed for generating audio descriptions for movies. Audio description (AD), also referred to as a video description, described video, or visual description, is a form of narration used to provide information surrounding key visual elements in a media work for the benefit of blind and visually impaired consumers. Audio descriptions are typically placed within a gap in dialogue or important sound.
Here are some examples of audio descriptions that were generated for a different movie clip:

- Example 1: Sitting on a nearby bench, the nerd looks over.
- Example 2: Now students spar in a hand-to-hand combat class.
- Example 3: He slams him down on the mat and raises his fist.
- Example 4: She stares distantly, then faces Andy.
- Example 5: A young girl cycles through the city at night.

The `System Prompt` establishes the model's role as an audio description generator and defines the output format requirements, ensuring consistent structure across all generated descriptions. The `User Prompt` provides the necessary context including video metadata, temporal boundaries from the localization stage, and ground-truth captions to guide the generation process. This two-stage prompting strategy enables the model to learn the mapping from localized temporal segments to natural language audio descriptions.

User Prompt

<video>

Watch the following movie clip and generate an Audio Description for it. The Audio Description should only consist of one sentence. Make sure you do not hallucinate information and only describe what you can see in the movie clip.
Requirements:

- Describe only the most important visible information in the clip.
- Focus on the visible people or subjects, their actions, and important interactions.
- Use present tense.
- Do not name characters unless their identity is explicitly provided in the input or visually confirmed. Instead use neutral words such as "A man", "A woman", "He", "She" etc. Assistant target the GT AD sentence, e.g.: Sitting on a nearby bench, the nerd looks over.

G Additional Qualitative Results

Figures 14 and 15 illustrate our model's localization and generation performance on CMD-AD test clips, while Figures 16 and 17 show results on LongLSMDC. Each example shows: (top) sampled video frames from the continuous clip, (middle) mel-spectrogram visualization where bright regions indicate audio activity and dark regions indicate silence periods suitable for AD insertion, and (Bottom) temporal alignment where black bars represent ground-truth AD segments and blue bars show our predicted segments.

CMD-AD Results: Learning Selective Silence Awareness. Figure 15 demonstrates that our proposed localization approach learns not every silence period requires an AD. In the top example, the mel-spectrogram shows multiple silence gaps (dark regions) throughout the 25-second clip, yet our model correctly predicts only 4 AD segments that align with ground truth. This selectivity reveals that the model distinguishes between acoustically silent moments that are narratively important versus those that are not. The audio-visual fusion enables this discrimination while audio detects silence periods, the visual stream determines whether the corresponding visual content is salient enough to describe. For instance, brief pauses during continuous action may be acoustically silent but visually redundant, requiring no narration. The predicted segments (blue bars) consistently align with both mel-spectrogram silence regions (dark bands) and ground-truth annotations (black bars), confirming that our approach successfully learns the joint constraint of acoustic availability and visual salience.

LongLSMDC Results: Dense, Continuous Annotations. Figures 16 and 17 showcase performance on densely annotated sequences where ADs occur in rapid succession, reflecting real-world untrimmed movie scenarios. Figure 17 is particularly challenging: 4-5 adjacent ground-truth AD segments occur within a 20-second window (timestamps 210-260s), with minimal gaps between them. This represents the high-density scenario characteristic of LongLSMDC, where ADs must be inserted continuously during extended dialogue-free sequences. Our model maintains temporal precision despite this tight spacing, with predicted segments (blue bars) closely tracking ground truth across the rapid sequence. The mel-spectrogram reveals sustained low-activity regions (dark purple bands) spanning multiple seconds, and our proposed localization approach successfully partitions these extended silence periods into multiple discrete AD segments based on visual changes. This demonstrates robust generalization to the challenging case of continuous narrative coverage where the model must identify not just isolated silence gaps, but also how to temporally segment extended silent periods into appropriate segments. Some prediction errors occur due to the extreme difficulty of these scenarios, highlighting the challenges of dense, real-world movie description tasks.

H Artifact License

The datasets and models employed in this work are subject to the following licenses. The Qwen2.5 and Qwen3 model families are released under the Apache 2.0 License, which permits free use, modification, and distribution for both research and commercial purposes. The Condensed Movies Dataset (CMD) is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which allows unrestricted use provided appropriate credit is given. The Large Scale Movie Description Challenge (LSMDC) dataset is made available under a restricted-access agreement through the Max Planck Institute for Informatics (MPII), permitting use exclusively for non-commercial academic research. The MAD-Eval (Movie Audio Descriptions) dataset is governed by a Non-Disclosure Agreement (NDA) that limits its use to academic research purposes only. All resources are therefore used in full compliance with their respective licensing terms within the scope of this academic work.

I Model size and budget

Stage 1. The audio-visual dual-head localizer has 6.64M parameters in total, shared across all three variants (*video-only*, *audio-only*, *full*); mode flags disable compute paths but do not change parameter count.

Stage 2. We use Qwen2.5-VL-7B or Qwen3-VL-8B as the backbone, adapted with LoRA (r=16, α=32, dropout 0.05) on the language model and multimodal projection; the vision encoder is frozen.

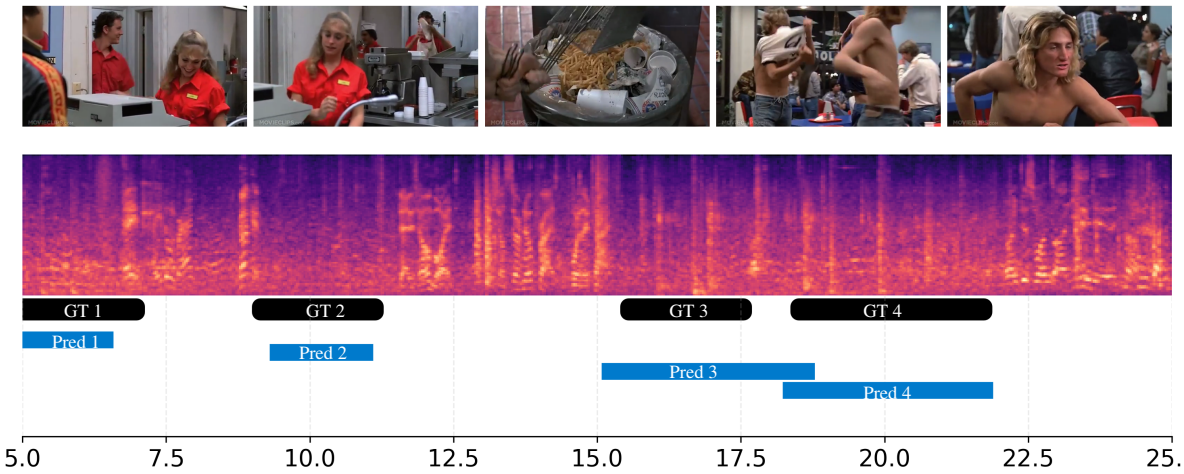
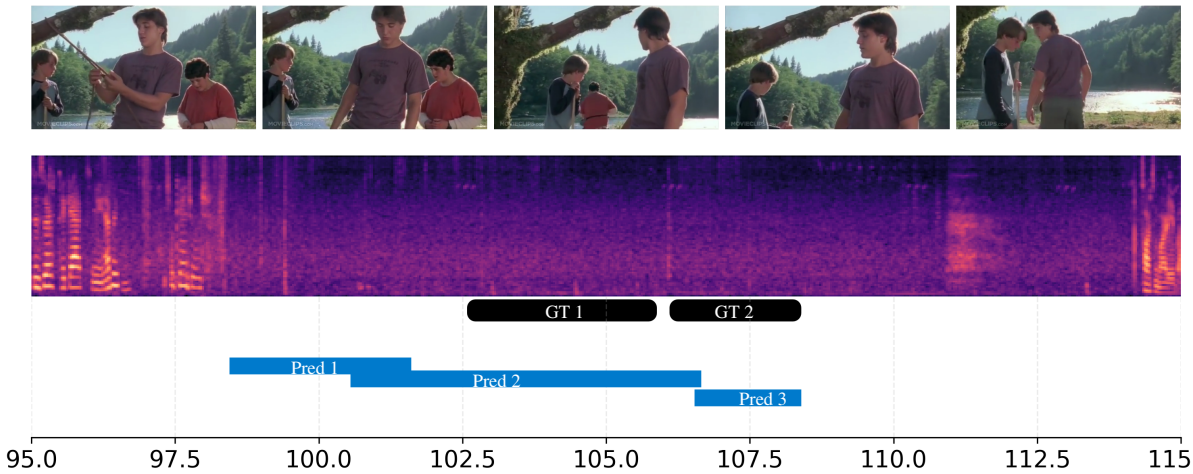


Figure 14: **CMD-AD qualitative examples on curated short clips.** Our model demonstrates effective localization and generation on moderately dense sequences. (Top) Sampled video frames, (Middle) mel-spectrogram showing audio activity (bright regions) and silence periods (dark regions), (Bottom) temporal alignment with ground-truth segments (black bars) and predicted segments (blue bars). The model successfully identifies appropriate silence windows and generates contextually relevant descriptions. GT: Ground Truth, Pred: Prediction.

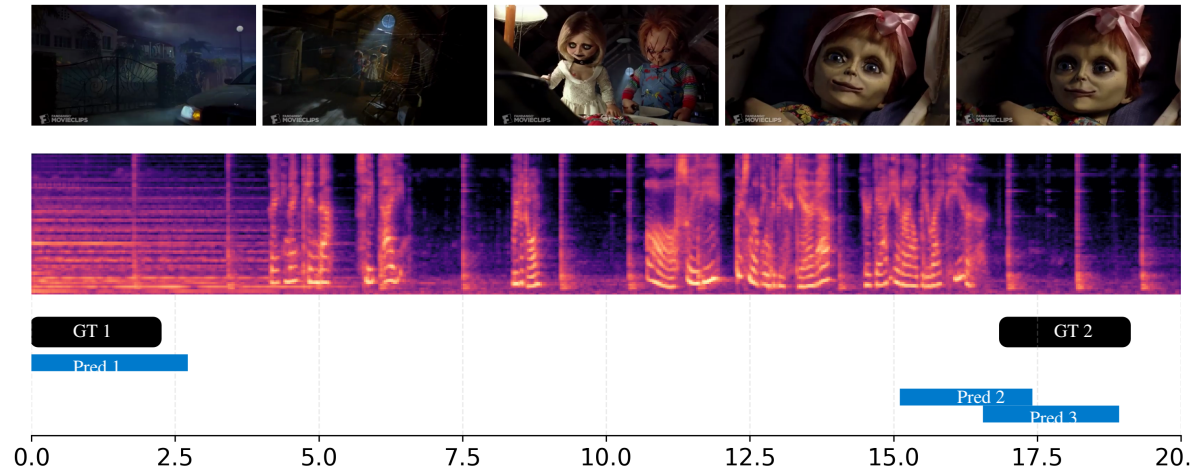


GT 1: George packs up his Swiss Army knife and plods away.

Pred 2: George folds up his knife and slowly walks away.

GT 2: Sam and Rocky watch him walking away.

Pred 3: He stands on one leg, then the other.



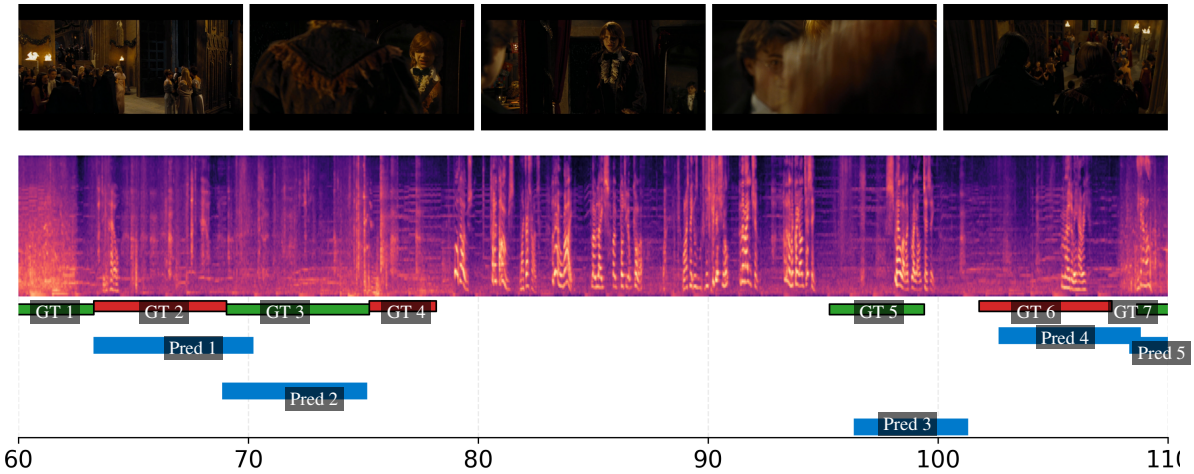
GT 1: Tiffany reaches for the lamp.

Pred 1: The light shines through the window above the attic door.

GT 2: Glen, Glenda stares at Chucky.

Pred 2: Tiffany and Chucky look at each other.

Figure 15: **CMD-AD example demonstrating selective silence awareness.** The model learns that not every acoustic silence requires an audio description. In the top example, multiple silence gaps appear in the mel-spectrogram throughout the 25-second clip, yet the model correctly predicts only segments corresponding to visually salient events that align with ground truth. This demonstrates that audio-visual fusion enables discrimination between acoustically silent moments that are narratively important versus those that are not. The visual stream determines content salience while audio detects insertion opportunities. GT: Ground Truth, Pred: Prediction.



GT 1: The girls wear long gowns, the boys dress robes.

GT 2: In his room, Ron adjusts his tatty robe.

Pred 1: HARRY is in the Great Hall, at the Gryffindor table.

GT 3: Ron stares at himself in a mirror.

Pred 2: The queen turns to her son.

GT 4: Harry comes in wearing a smart black robe and a white bow tie.

GT 5: Harry takes a deep breath and sinks down after it.

Pred 3: They look at each other.

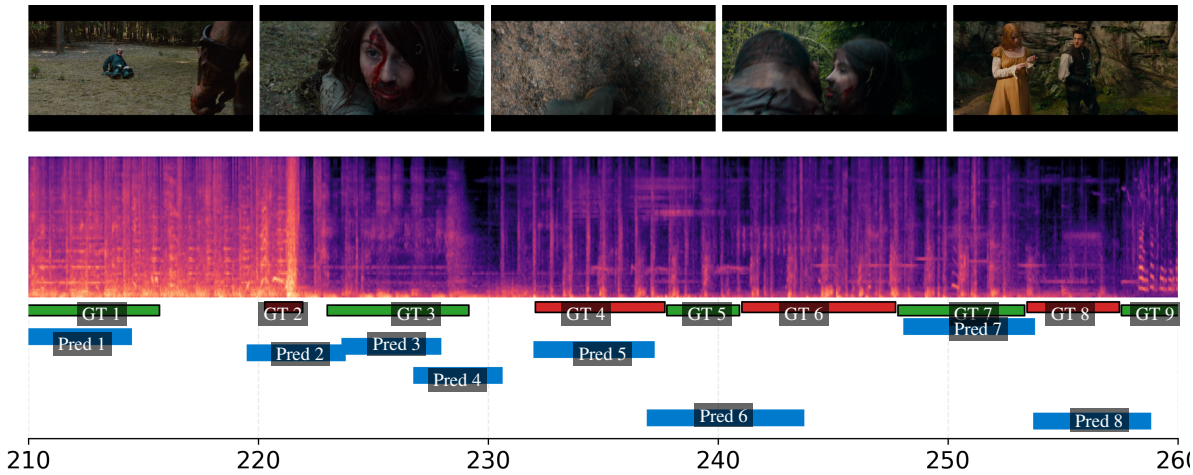
GT 6: Moody searches the shelves.

Pred 4: They are all wearing costumes.

GT 7: Snowflakes whirl outside the boys' bedroom.

Pred 5: They both wear bow ties as it snows.

Figure 16: **LongLSMDC example showing dense annotations in real-world scenarios.** LongLSMDC presents significantly more challenging scenarios with high annotation density and longer, untrimmed clips. This example shows 7 ground-truth segments within a 50-second window, demonstrating rapid succession of events typical of continuous movie sequences. The model maintains temporal precision despite tight spacing, with predicted segments closely tracking ground truth. The mel-spectrogram reveals sustained silence periods that must be partitioned into multiple discrete AD opportunities based on visual content changes, demonstrating the model's ability to handle continuous narrative coverage. GT: Ground Truth, Pred: Prediction.



GT 1: Mina stands with her wand ready after assaulting Muriel with a spell.

Pred 1: A man with long hair and a beard, dressed in leather, and carrying a spear, lifts a woman into his arms.

GT 2: The troll stomps on his head, his brains and blood splattering on the ground.

Pred 2:

GT 3: Mina shatters a tree behind Muriel.

Pred 3: The man stands in the middle of a clearing.

Pred 4: The giant turns and sees Jane.

GT 4: Gretel awakens to the gentle rocking motion of being carried.

Pred 5: The light from the flashlight beam moves about the ground.

GT 6: Gretel turns her head to find the side of the troll's head.

GT 5: Ben watches Gretel fall.

Pred 6: John reaches out to grab the giant's foot.

GT 7: Gretel closes her eyes and leans her head to rest against the giant's shoulder.

Pred 7: A man with long hair and a beard, dressed in leather, and carrying a spear, lifts a woman into his arms.

GT 8: They kiss.

Pred 8: Muriel and her other witch, one with red hair, hover menacingly.

Figure 17: **LongLSMDC example showing dense annotation and real-world annotation scenarios.** This sequence contains 9 ground-truth segments within a 50-second window (4-5 adjacent segments in rapid succession around timestamps 210-260s), representing the most challenging high-density scenarios in real-world untrimmed movies. The model must handle: (1) minimal gaps between consecutive ADs, (2) complex visual scenes with multiple characters and actions, (3) extended dialogue-free sequences requiring continuous narrative coverage. Despite these challenges, the localizer successfully partitions extended silence periods into appropriate description units. Some prediction errors occur due to the extreme difficulty of this scenario, highlighting areas for future improvement. GT: Ground Truth, Pred: Prediction.