STRUCTURED POSTERIOR DECODING OVER CONTEXTUALIZED SPAN LATTICES FOR VIDEO MOMENT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

A central challenge in video moment retrieval is assigning reliable relevance scores among many plausible temporal spans. Recent work in audio moment retrieval showed that exact segment posteriors, even with simple boundary-based potentials, can provide an effective structured alternative to proposal confidence. We observe, however, that when this formulation is transferred directly to video, its gains diminish as more target moments must be retrieved. We attribute this to the different nature of temporal evidence in video: relevant visual semantics are often distributed throughout a moment, while boundary transitions tend to be less distinct. To address this, we contextualize the two-dimensional span lattice using query-conditioned local 2D convolutions, allowing each candidate to aggregate span-interior semantics and information from neighboring temporal hypotheses before structured inference. We then compute exact segment marginals and use them to form a posterior-weighted moment prototype, whose similarity to each candidate iteratively refines the span potentials. Under matched feature settings, our approach substantially improves multi-moment retrieval over prior methods on QV-M^2, and with InternVideo2-6B features it establishes a new state of the art on this benchmark. The gains carry over to QVHighlights and Charades-STA, while a cross-modal analysis on audio suggests that the benefit of span-lattice contextualization depends on modality-specific temporal evidence.

1 INTRODUCTION

Video moment retrieval (VMR) aims to localize temporal segments in an untrimmed video that correspond to a natural-language query (Anne Hendricks et al., 2017; Lei et al., 2021). While conventional VMR is often evaluated by identifying a highly relevant moment for each query, recent settings increasingly require retrieving multiple disjoint occurrences that satisfy the same query (Huang et al., 2024; Qin et al., 2025). This multi-moment retrieval (MMR) setting changes the nature of the problem: rather than selecting a single best span, the model must identify and rank several distinct relevant moments while suppressing redundant alternatives around them (Cao et al., 2025).

Accordingly, we view MMR as a structured ranking problem over a dense set of correlated temporal spans. Existing VMR models, however, are not explicitly designed around this requirement (Zhang et al., 2023). Methods that generate a finite set of moment candidates, including DEtection TRansformer (DETR)-based approaches, rank the resulting spans using proposal-level confidence scores, tying retrieval coverage to the candidate-generation mechanism itself (Moon et al., 2023; 2026). Other methods score a dense set of start–end combinations, alleviating this coverage constraint and even modeling interactions among neighboring candidates (Zhang et al., 2020; 2022). However, their final relevance scores are still assigned to individual candidates rather than jointly normalized over complete temporal explanations.

A natural way to introduce such competition is to model complete temporal segmentations rather than scoring candidate spans independently (Sarawagi & Cohen, 2004). Under a segmental model, each possible segmentation jointly explains the video timeline as a sequence of relevant moments and background intervals, and the relevance of a candidate span can be measured by its marginal probability of appearing across all possible segmentations (Zhuo et al., 2016). This exact segment

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

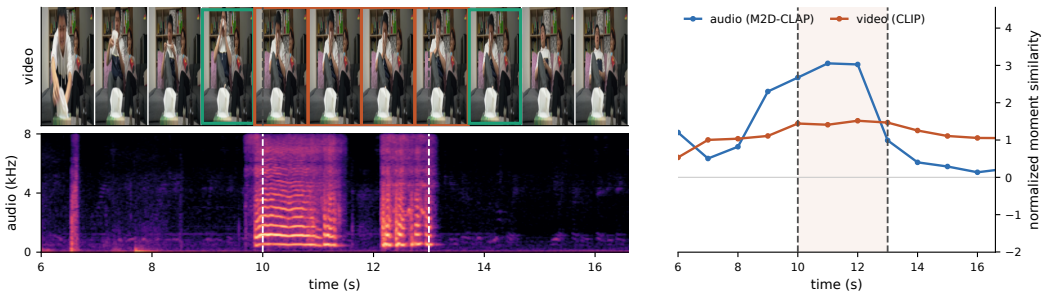


Figure 1: Illustrative audio–video evidence for the query "someone blows their nose loudly." Left: video frames and the corresponding audio spectrogram; right: normalized moment similarity from CLIP (Radford et al., 2021) and M2D-CLAP (Niizumi et al., 2025). Orange marks the annotated target moment, green marks visually similar neighboring frames outside the target.

posterior provides a globally normalized score over competing temporal explanations, while allowing multiple disjoint moments to coexist within the same structured hypothesis (Lafferty et al., 2001). Recent work in audio moment retrieval showed that such segment marginals can serve as an effective alternative to proposal confidence, even when the segment potentials are defined using simple boundary-based evidence (Choi et al., 2026). This makes segmental posterior decoding a natural starting point for MMR.

However, the benefit of this formulation diminishes in video as the number of target moments grows. In audio, boundary-local evidence can be highly informative, as onset and offset cues are explicitly exploited for temporal sound-event localization (Song, 2024; Choi et al., 2026). In contrast, visual transitions around event boundaries are often gradual or ambiguous (Su et al., 2019; Shi et al., 2023; Wei et al., 2025; Liu et al., 2025). As illustrated in Fig. 1, for the query "someone blows their nose loudly," the video remains similar across the annotated boundaries, whereas the corresponding acoustic event is more distinct. Structured decoding in video therefore requires richer span potentials that capture span-interior semantics and distinguish among neighboring temporal hypotheses.

To address this limitation, we draw on the 2D temporal-map formulation of 2D-TAN (Zhang et al., 2020; 2022), representing valid temporal spans on a two-dimensional start–end lattice. Rather than using this lattice only to score candidate moments, we contextualize the span space before structured inference. Specifically, we encode span-interior visual evidence for each candidate and propagate information across neighboring hypotheses using query-conditioned local 2D convolutions, producing context-aware segment potentials for global normalization. We further make the structured posterior part of the representation process itself. After computing exact segment marginals, we aggregate candidate representations into a posterior-weighted moment prototype and use its similarity to each candidate to iteratively refine the segment potentials. In this way, span-lattice contextualization captures local relations among candidate moments, while posterior-conditioned refinement feeds globally normalized retrieval beliefs back into their potentials.

Across experiments, our approach improves over prior SOTA methods on QV-M^2 (Cao et al., 2025) under matched feature settings, with particularly strong gains in multi-moment retrieval. When combined with InternVideo2-6B features (Wang et al., 2024), it further establishes a new state of the art on QV-M^2. These improvements also transfer to additional VMR benchmarks, showing that the proposed modeling is not specific to a single dataset. Cross-modal analysis suggests that the effectiveness of span contextualization depends on modality-specific temporal evidence, motivating potential designs that are adapted to each modality. Our contributions are summarized as follows:

- We propose structured posterior decoding over contextualized span lattices for VMR, combining query-conditioned 2D span contextualization with posterior-conditioned iterative potential refinement.
- Our method achieves consistent improvements across VMR benchmarks, including particularly strong gains in multi-moment retrieval and new state-of-the-art performance on QV-M^2 with InternVideo2-6B features.

- Cross-modal analysis suggests that the effectiveness of span-lattice contextualization depends on modality-specific temporal evidence, motivating modality-adapted segment potentials.

2 Related Work

Proposal-based video moment retrieval and multi-moment retrieval. Recent VMR methods increasingly adopt transformer-based set prediction to directly generate temporal moments from query-conditioned video representations. Moment-DETR formulates moment retrieval as a direct set prediction problem and predicts a fixed set of moment coordinates and confidence scores using a transformer encoder–decoder (Lei et al., 2021). QD-DETR further strengthens query conditioning by explicitly injecting textual context into the video representation through cross-attention, improving both moment retrieval and highlight detection (Moon et al., 2023). Subsequent methods have continued to improve cross-modal representation (Moon et al., 2026), temporal modeling (Jing et al., 2023; Zhou et al., 2025), and proposal scoring (Gordeev et al., 2026) while retaining candidate-based decoding. More recently, the QV-M^2 benchmark explicitly introduced multi-moment retrieval, where a single query may correspond to several disjoint temporal occurrences (Cao et al., 2025). Its accompanying FlashMMR framework refines candidate boundaries and re-evaluates the resulting moments through multi-moment post-verification, improving the selection of multiple relevant segments. These approaches demonstrate substantial progress in candidate generation and verification, whereas our work focuses on a complementary question: how relevance should be assigned when many plausible spans compete to explain the same video. Rather than relying on confidence scores attached to a finite set of proposals, we rank spans using their posterior probabilities under a globally normalized distribution over complete temporal segmentations.

Dense span representations and structured temporal modeling. An alternative line of VMR research represents temporal candidates densely in a two-dimensional span space. 2D-TAN arranges candidate moments on a temporal map indexed by their boundaries and applies convolutional operators to capture relations among adjacent candidates (Zhang et al., 2020); MS-2D-TAN extends this idea to multiple temporal scales to model contextual relations among moments of different durations (Zhang et al., 2022). These methods show that interactions among neighboring span hypotheses provide useful context beyond independently encoded candidates. Our method adopts this dense span-lattice perspective but uses the contextualized lattice to construct potentials for structured inference rather than directly treating map scores as final retrieval scores. Structured prediction offers a complementary mechanism for modeling competition among temporal hypotheses. Conditional random fields define globally normalized distributions over structured outputs (Lafferty et al., 2001), while semi-Markov CRFs extend this formulation to segment-level variables, allowing features and potentials to be defined over complete spans and retaining exact dynamic-programming inference (Sarawagi & Cohen, 2004). Recent work in audio moment retrieval applied this principle to temporal retrieval by ranking candidate moments with exact segment marginal posteriors instead of DETR proposal confidence (Choi et al., 2026). We build on this posterior-decoding principle, but address the video-specific challenge of constructing informative segment potentials: we contextualize span-interior visual evidence and neighboring hypotheses on the 2D lattice, and further refine the potentials using feedback from the globally normalized posterior.

3 Methodology

3.1 Structured Posterior Formulation

Figure 2 provides an overview of the proposed framework. Starting from query-conditioned video representations, the model constructs contextualized span potentials over a 2-D lattice and iteratively refines them using feedback from exact segment posteriors before final posterior decoding.

The proposed method builds on the segmental posterior decoding framework of Choi et al. (2026). We summarize below the elements necessary for the video-specific extensions. Given a video–query pair, the QD-DETR encoder produces a query-conditioned video memory $\mathbf{H} = (\mathbf{h}_t)_{t=1}^{L}$. A lightweight frame-level head predicts onset, offset, and background scores,

$$(o_t, r_t, b_t) = f_{\text{frame}}(\text{LN}(\mathbf{h}_t)), \qquad \psi_{\text{base}}(s, e) = o_s + r_{e-1} + b_{\text{fg}}, \qquad \phi_{\text{bg}}(t) = b_t. \tag{1}$$

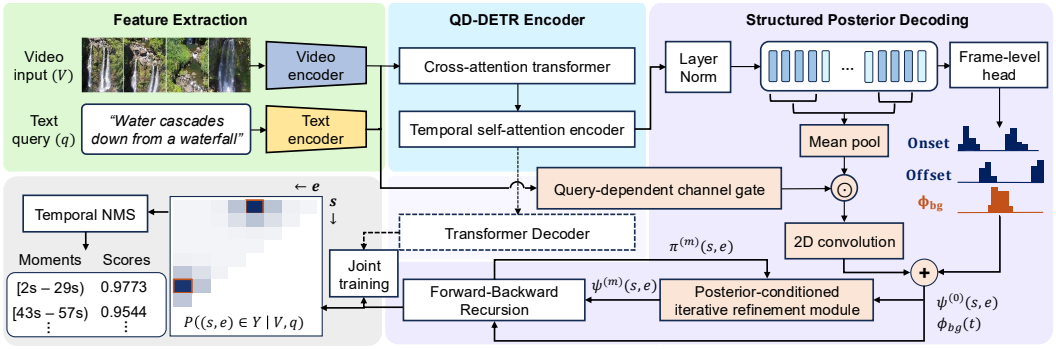


Figure 2: Overview of the proposed VMR framework. Dashed components are used only during training, and orange components indicate the proposed span-lattice contextualization and posterior-conditioned refinement modules.

Here, $\psi_{\text{base}}(s, e)$ is the potential of a foreground segment $[s, e)$ and $\phi_{\text{bg}}(t)$ is the background emission at time t. We retain the boundary-only potential of the base structured decoder and introduce span-interior evidence explicitly through the video-specific lattice contextualizer described next.

Let Y denote a temporal segmentation consisting of non-overlapping foreground segments, with uncovered frames assigned to background. Its score and globally normalized distribution are

$$S(Y) = \sum_{(s,e)\in Y} \psi(s, e) + \sum_{t \notin \cup_{(s,e)\in Y}[s,e)} \phi_{\text{bg}}(t), \qquad P(Y \mid \mathbf{V}, \mathbf{Q}) = \frac{\exp S(Y)}{Z}, \tag{2}$$

where $Z = \sum_{\widetilde{Y}\in\mathcal{Y}} \exp S(\widetilde{Y})$.

Segmental forward–backward gives the exact marginal probability of every valid candidate span,

$$\pi_{s,e} \triangleq P((s, e) \in Y \mid \mathbf{V}, \mathbf{Q}) = \exp(\alpha_s + \psi(s, e) + \beta_e - \log Z)\,, \tag{3}$$

where α and β denote the forward and backward quantities.

The structured branch is trained with the per-frame-normalized negative log-likelihood

$$\mathcal{L}_{\text{NLL}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\log Z_b - S(Y_b^\star)}{L_b}, \tag{4}$$

together with the minimum IoU risk

$$\mathcal{L}_{\text{risk}} = 1 - \frac{\sum_{(s,e)} \pi_{s,e}\, u_{s,e}}{\sum_{(s,e)} \pi_{s,e}}, \qquad u_{s,e} = \max_{(s^\star,e^\star)\in Y^\star} \text{IoU}([s, e), [s^\star, e^\star))\,. \tag{5}$$

The overall objective is $\mathcal{L} = \mathcal{L}_{\text{DETR}} + 0.5\,\mathcal{L}_{\text{NLL}} + \mathcal{L}_{\text{risk}}$, where $\mathcal{L}_{\text{DETR}}$ denotes the DETR-branch losses (Table 10).

3.2 Query-Conditioned Span-Lattice Contextualization

Although global normalization introduces competition among temporal explanations, the base potential in Eq. 1 depends primarily on boundary evidence. For video, where relevant semantics are often distributed throughout a moment, we therefore construct a dense representation over valid temporal spans and contextualize this representation before structured inference.

For each video token, we first project the query-conditioned memory as

$$\mathbf{z}_t = \mathbf{W}_{\text{in}}\text{LN}(\mathbf{h}_t), \qquad \mathbf{g}_q = \sigma(\mathbf{W}_{\text{txt}}\bar{\mathbf{q}})\,, \tag{6}$$

where $\bar{\mathbf{q}}$ is the pooled query representation and $\mathbf{g}_q$ is a query-dependent channel gate. The representation of a candidate span $[s, e)$ is then

$$\mathbf{F}_{e,d} = \left(\frac{1}{d}\sum_{t=e-d}^{e-1} \mathbf{z}_t\right) \odot \mathbf{g}_q, \qquad d = e - s \tag{7}$$

Thus, each lattice cell explicitly encodes the visual content within the corresponding temporal span, while remaining conditioned on the text query. Span means are computed using prefix sums, avoiding repeated aggregation over overlapping intervals.

We next contextualize the lattice with a stack of local two-dimensional convolutions,

$$\mathbf{M} = f_{\mathrm{2D}}(\mathbf{F}), \qquad \psi^{(0)}(s,e) = \psi_{\mathrm{base}}(s,e) + \lambda_{\mathrm{2D}} M_{e,e-s}, \tag{8}$$

where $M_{e,e-s}$ is a scalar contextual score for the candidate $[s,e)$. The convolutional receptive field spans both temporal position and duration, allowing each candidate to exchange information with neighboring hypotheses that have nearby boundaries or extents.

Importantly, the contextualized map is not used as an independently normalized retrieval score. Instead, it modifies the segment potentials before the global normalization in Eq. 2. This combines span-interior visual evidence and local interactions among candidate moments with the structured competition induced by the segmental posterior. The final projection of f_{2D} is zero-initialized, so the model initially reduces exactly to the base structured decoder.

3.3 Posterior-Conditioned Iterative Refinement

The 2-D lattice in Sec. 3.2 captures interactions among locally neighboring hypotheses, but does not directly exploit the global retrieval belief induced by structured normalization. We therefore feed the segment posterior back into the potential construction through an iterative refinement procedure.

For each candidate span $[s,e)$, we construct an ℓ_2-normalized span representation

$$\mathbf{c}_{s,e} = \mathrm{norm}\left(\frac{1}{e-s}\sum_{t=s}^{e-1} \mathbf{W}_{\mathrm{ref}}\mathrm{LN}(\mathbf{h}_t)\right), \tag{9}$$

where $\mathrm{norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. Starting from the contextualized potential $\psi^{(0)}$ in Eq. 8, iteration m first recomputes the exact segment marginals

$$\pi^{(m)}_{s,e} = P_{\psi^{(m-1)}}((s,e) \in Y \mid \mathbf{V}, \mathbf{Q}). \tag{10}$$

These marginals are normalized across candidate spans to form a posterior-weighted prototype,

$$w^{(m)}_{s,e} = \frac{\mathrm{sg}\left(\pi^{(m)}_{s,e}\right)}{\sum_{(i,j)\in\mathcal{C}_{\mathrm{ref}}} \mathrm{sg}\left(\pi^{(m)}_{i,j}\right)}, \qquad \boldsymbol{\mu}^{(m)} = \mathrm{norm}\left(\sum_{(s,e)\in\mathcal{C}_{\mathrm{ref}}} w^{(m)}_{s,e}\mathbf{c}_{s,e}\right), \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\mathcal{C}_{\mathrm{ref}}$ is the set of spans considered by the refinement module.

The resulting prototype summarizes the visual characteristics of spans currently supported by the globally normalized posterior. We refine each candidate potential according to its similarity to this prototype:

$$\psi^{(m)}(s,e) = \psi^{(0)}(s,e) + \lambda_{\mathrm{ref}}\, \mathbf{c}_{s,e}^{\top}\boldsymbol{\mu}^{(m)}, \qquad (s,e) \in \mathcal{C}_{\mathrm{ref}}. \tag{12}$$

Candidates consistent with the posterior-supported prototype therefore receive increased potential, while inconsistent candidates are relatively suppressed. Repeating this procedure allows the globally normalized retrieval belief to iteratively influence the local span potentials.

Each iteration performs forward–backward inference for its current fixed potential table. Consequently, the final decoding remains the exact segment posterior under the final refined potentials, rather than an approximate inference procedure for an implicit higher-order model. The refinement scale λ_{ref} is initialized to zero, so the module initially reduces exactly to the contextualized model of Sec. 3.2. We use two refinement iterations in all main experiments, and analyze the iteration count in Appendix A.2.4.

4 Experimental Settings

4.1 Datasets and Evaluation

We primarily evaluate on QV-M^2 (Cao et al., 2025), a multi-moment re-annotation of QVHighlights designed to retrieve multiple disjoint moments for a single query. Its test split is substantially more

Table 1: Comparison with state-of-the-art methods on the QV-M^2 test set. Results for prior methods are taken from Cao et al. (2025). Our results in this table correspond to the best-performing model. The best results are highlighted in **bold**, and the second-best are underlined.

Method	mAP				mIoU@k			mR@k		
	G-mAP	@1tgt	@2tgt	@3+tgt	@1	@2	@3	@1	@2	@3
QD-DETR	28.95	39.69	37.26	18.30	50.46	46.79	40.50	42.35	40.58	36.05
TR-DETR	31.23	44.12	39.17	19.64	55.13	48.13	42.52	47.21	41.24	35.82
CG-DETR	28.87	43.74	32.35	18.44	52.01	47.98	43.27	43.26	40.80	36.69
FlashVTG	32.14	47.16	39.48	20.19	54.49	47.85	40.92	46.64	41.30	35.94
FlashMMR	35.14	52.59	42.52	22.89	56.29	49.64	42.92	48.81	44.33	38.50
Ours (CLIP+SF)	36.89	49.28	**48.09**	**28.40**	56.24	49.68	42.10	50.20	**45.32**	38.90
Ours (IV2)	**39.18**	**62.39**	43.05	26.57	**60.76**	**50.92**	**43.29**	**53.03**	44.40	**39.16**

multi-moment-heavy than the training and validation splits, with 3.31 annotated moments per query on average. To assess generality beyond this setting, we additionally evaluate on QVHighlights (Lei et al., 2021) and Charades-STA (Gao et al., 2017), covering conventional VMR settings with fewer target moments per query. We further use CASTELLA (Munakata et al., 2026) for the cross-modal transfer analysis in Sec. 5.4. Detailed dataset statistics are provided in Appendix A.1.1.

We follow the evaluation protocol of Cao et al. (2025). G-mAP is the mean query-level average precision over tIoU thresholds from 0.5 to 0.95. $\mathrm{mAP@1tgt}$ and $\mathrm{mAP@2tgt}$ evaluate queries with exactly one and two ground-truth moments, respectively, while $\mathrm{mAP@3{+}tgt}$ averages performance across ground-truth-count groups with three or more targets. $\mathrm{mR@}k$ measures the fraction of targets (up to k per query) recovered by the top-k predictions, averaged over multiple tIoU thresholds, and $\mathrm{mIoU@}k$ measures the localization quality of their one-to-one matches. We additionally report the standard VMR metrics R1@0.5 and R1@0.7. Unless otherwise stated, the structured channel ranks the top 50 exact segment marginals, applies temporal non-maximum suppression (NMS) at IoU 0.7, and retains the top 10 predictions. Formal metric definitions and evaluation details are given in Appendix A.1.2.

4.2 Implementation Details

For VMR, we use a QD-DETR-based architecture (Moon et al., 2023) with frozen pretrained video and text features[1]. The main setting uses CLIP ViT-B/32 (Radford et al., 2021) and SlowFast (Feichtenhofer et al., 2019) video features together with CLIP text features. The stronger feature setting replaces them with InternVideo2-6B (Wang et al., 2024) and LLaMA (Touvron et al., 2023) features, respectively. For the cross-modal analysis on CASTELLA, we retain the underlying AMR system and its training protocol (Choi et al., 2026) unchanged, using frozen M2D-CLAP (Niizumi et al., 2025) audio and text features with Clotho-Moment pretraining followed by CASTELLA finetuning. All remaining architectural and optimization details are provided in Appendix A.2.

5 Results & Discussions

5.1 Comparison with Existing Methods

Table 1 compares our method with existing approaches on the QV-M^2 test set. Under the matched CLIP+SlowFast setting, our method achieves 36.89 G-mAP and shows particularly strong performance on multi-moment queries. While its single-target performance is comparable to prior methods, the gains become more pronounced as multiple relevant moments must be retrieved and ranked. This pattern is consistent with the intended role of structured segmental modeling in resolving competition among plausible temporal candidates.

[1]Upon acceptance, we will release the source code, configuration files, and evaluation scripts on GitHub to facilitate reproducibility.

Table 2: Ablation and generalization results. All results are averages over three independently trained seeds. Base (DETR) uses conventional decoder confidence scoring. Base (SPD) uses segmental posterior decoding from Choi et al. (2026), 2DCon additionally contextualizes the span lattice, ITRef applies posterior-conditioned iterative refinement, and Full combines 2DCon and ITRef.

Method	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt	mIoU@1	mIoU@2	mIoU@3	mR@1	mR@2	mR@3
(a) QV-M^2 test — CLIP+SlowFast												
Base (DETR)	49.74	37.36	25.48	35.85	34.24	15.14	47.79	43.43	38.08	40.03	36.11	30.91
Base (local-ψ)	37.36	23.53	16.19	32.67	21.40	4.55	38.99	25.10	18.19	29.20	19.64	14.70
Base (SPD)	55.93	43.45	32.42	47.55	40.85	20.98	53.57	45.63	38.34	46.86	40.78	34.65
2DCon	53.56	42.72	34.17	48.49	42.39	24.98	53.50	46.08	39.00	46.34	41.03	35.68
ITRef	56.25	**44.99**	33.69	**50.79**	42.88	21.01	**54.76**	46.47	38.11	**48.31**	41.66	34.83
Full (local-ψ)	52.53	40.87	28.03	40.43	38.68	19.40	51.47	42.58	34.42	44.29	37.12	30.88
Full (Ours)	**56.45**	44.17	**34.95**	46.43	**44.81**	**26.86**	54.25	**47.51**	**40.95**	47.70	**43.17**	**37.61**
(b) QV-M^2 test — InternVideo2-6B + LLaMA												
Base (DETR)	56.14	40.35	28.83	45.08	34.22	16.13	53.36	46.37	41.05	44.31	38.09	34.47
Base (SPD)	59.75	46.34	34.51	56.64	41.16	19.51	57.45	47.77	39.60	49.42	41.85	35.84
Full	**62.02**	**48.61**	**38.43**	**59.18**	**44.60**	**26.62**	**59.88**	**49.90**	**42.52**	**52.23**	**44.14**	**38.54**
(c) QVHighlights test												
Base (DETR)	60.29	43.92	38.71	47.69	25.42	7.71	55.46	39.18	33.98	47.50	31.36	26.59
Base (SPD)	61.28	45.78	42.97	51.81	**30.47**	10.52	57.77	39.11	31.92	50.23	33.96	27.80
Full	**62.30**	**47.15**	**44.70**	**53.91**	30.46	**13.20**	**58.98**	**39.92**	**34.25**	**51.39**	**34.89**	**30.29**
(d) Charades-STA test												
Base (DETR)	57.11	35.18	34.78	34.78	–	–	49.31	–	–	40.75	–	–
Base (SPD)	56.77	33.68	37.37	37.37	–	–	50.12	–	–	40.65	–	–
Full	**57.31**	**36.61**	**39.50**	**39.50**	–	–	**50.17**	–	–	**41.78**	–	–

With stronger InternVideo2-6B and LLaMA features, G-mAP further increases to 39.18. Interestingly, the improvement is concentrated more strongly in single-target retrieval, while multi-moment performance remains close to that obtained with CLIP+SlowFast features. This suggests that the stronger representations primarily improve single-target retrieval in this benchmark, whereas retrieving multiple relevant moments continues to depend on how candidate spans are structured and compared.

5.2 Ablation Study

We use the ablations to separate three questions: the role of global normalization, the contribution of each proposed module, and whether the gains persist across representations and datasets. Table 2 reports three-seed averages for the component ablation and in-domain evaluation on additional VMR benchmarks. Full mean and standard deviation results are provided in the appendix.

Component contributions. On QV-M^2, Base (SPD) improves over conventional DETR confidence decoding, but its gain shrinks from 11.7 AP on single-target queries to 5.8 AP on queries with three or more targets. The two proposed extensions exhibit complementary effects. 2DCon mainly drives the gains on multi-moment queries, whereas ITRef is more effective for top-1 retrieval and queries with fewer targets. Their combination in Full provides the best overall trade-off, achieving the highest G-mAP and the strongest multi-target retrieval performance.

We additionally compare structured posterior decoding against a local ψ scorer without global normalization trained and decoded over the same span lattice. Although 2-D contextualization substantially strengthens the local scorer, Full (local-ψ) remains below its SPD counterpart, with the largest gap appearing on multi-moment retrieval. This indicates that contextualizing individual span scores and globally normalizing competing span hypotheses provide complementary benefits, with structured decoding becoming particularly important when multiple relevant moments must be ranked.

To test whether these gains merely compensate for limited visual representations, we further compare Base (DETR), Base (SPD), and Full using InternVideo2-6B video features with LLaMA text features. The improvements from SPD and the full model persist with the stronger representations, showing that structured span modeling provides gains beyond those obtained from stronger pretrained features alone.

Table 3: Effect of neighboring-span interactions induced by the 2-D convolution over the span lattice on QV-M^2. Results are three-seed means. Kernel sizes denote the 2-D convolutional kernel over the end-time $\times$ duration axes. The capacity-matched 1×1 model uses approximately the same number of total 2DCon parameters as the 3×3 model. All variants include ITRef.

Neighbor interaction	Kernel	2DCon Params	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt
None	1×1	20,962	54.70	43.14	32.11	46.85	42.00	20.91
None (capacity-matched)	1×1	54,334	56.04	**45.10**	32.95	**48.70**	43.44	19.76
Duration neighbors	1×3	29,154	55.42	43.34	32.23	43.85	41.10	21.86
End-time neighbors	3×1	29,154	54.28	42.83	33.12	44.85	43.03	24.58
End + duration neighbors	3×3	53,730	**56.45**	44.17	**34.95**	46.43	**44.81**	**26.86**

Table 4: Effect of neighboring-span interactions induced by the 2-D convolution over the span lattice on CASTELLA audio moment retrieval. Results are three-seed means on the CASTELLA test set, evaluated with the QV-M^2 multi-moment evaluator under the full-duration ($d_{\max} = 300$) setting. Kernel sizes denote the 2-D convolutional kernel over the end-time $\times$ duration axes.

Neighbor interaction	Kernel	2DCon Params	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt
Base (SPD)	–	–	52.36	**40.96**	**34.11**	42.09	35.89	14.79
None	1×1	20,962	51.62	40.09	33.98	**42.18**	**36.43**	14.73
Duration neighbors	1×3	29,154	**52.86**	40.54	34.07	42.17	35.42	**15.92**
End-time neighbors	3×1	29,154	51.72	39.77	33.47	41.48	35.83	14.80
End + duration neighbors	3×3	53,730	52.07	39.25	33.15	40.86	35.49	14.93

Evaluation across additional VMR benchmarks. We additionally train the same framework in-domain on QVHighlights and Charades-STA. The progression from DETR confidence decoding to SPD and Full is preserved on QVHighlights, including its multi-target retrieval metrics. The gains also transfer to Charades-STA, where each query contains only one annotated moment. Thus, while the largest improvements on QV-M^2 occur for multi-moment queries, the benefit of structured span modeling is not restricted to the multi-moment setting.

5.3 Analysis of Span-Lattice Context

Table 3 examines whether the gain from 2DCon comes from its additional capacity or from interactions among neighboring span hypotheses. A 1×1 convolution, which processes each span independently, does not improve G-mAP or multi-target mAP over ITRef alone (Table 2), even when its channel width is increased to match the parameter count of the 3×3 model. In contrast, expanding the receptive field along the end axis substantially improves retrieval for queries with multiple targets, and jointly contextualizing both lattice axes achieves the highest G-mAP and the strongest multi-target performance. These results indicate that the benefit of 2DCon comes primarily from relational context among neighboring spans rather than additional model capacity.

The same trend becomes more pronounced for queries containing more relevant moments. Within QV-M², lattice contextualization yields larger gains as the number of ground-truth moments increases. Together with the receptive-field ablation, this descriptive trend suggests that interactions among neighboring span hypotheses are particularly beneficial when more candidate moments compete for retrieval.

5.4 Cross-Modal Transfer to Audio and Boundary Evidence

Table 4 examines whether the span-lattice contextualization that helps video moment retrieval transfers to audio moment retrieval on CASTELLA. Unlike QV-M^2, where 2DCon yields clear gains on multi-target queries, no kernel yields a comparable gain on CASTELLA. Overall performance remains at or slightly below Base (SPD), with the full 3×3 neighborhood performing worst. Thus, the gain from neighboring-span interactions is not a generic property of temporal retrieval, but appears to be more specific to the video setting.

Figure 3 provides a complementary comparison of the learned local boundary evidence under the evaluated audio and video configurations. Around both the start and end boundaries, the audio

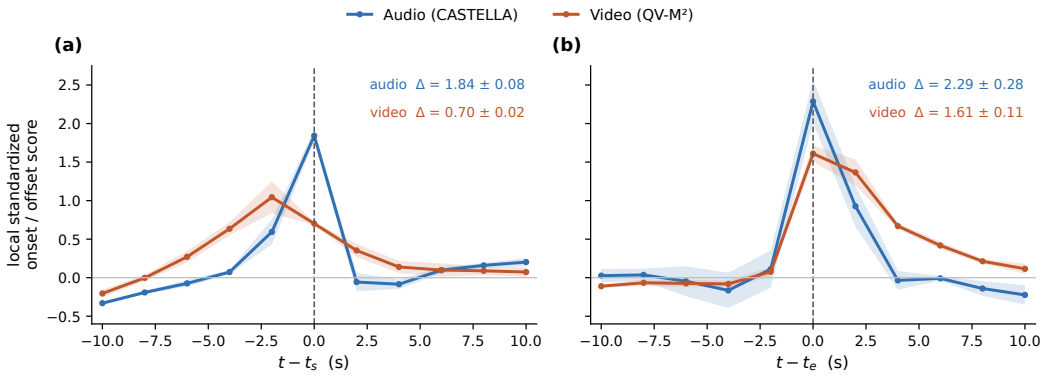


Figure 3: Query-conditioned local boundary evidence around ground-truth moment boundaries in audio and video. Panel (a) shows the mean standardized onset score aligned to the start time t_s, and panel (b) shows the mean standardized offset score aligned to the end time t_e. Curves are averaged over ground-truth moments, and shaded regions denote standard errors. Δ summarizes boundary sharpness as the difference between the score at the aligned boundary and nearby time steps.

model exhibits sharper and more concentrated onset/offset responses, whereas the video responses are broader and spread over a wider temporal region. Because the two modalities are evaluated with different benchmarks and feature encoders, this comparison should not be interpreted as isolating modality alone. Nevertheless, the observed pattern is consistent with the transfer results in Table 4, the evaluated audio setting relies more strongly on localized boundary evidence, whereas neighboring-span contextualization provides a larger benefit in the evaluated video setting. Overall, such contextualization appears most useful when local boundary evidence alone cannot separate competing candidates.

6 Limitations & Future Work

Our approach explicitly represents and contextualizes the two-dimensional span lattice, which enables interactions among competing temporal candidates but incurs quadratic memory and computation with respect to the temporal resolution. While this cost is manageable under the benchmark settings considered in this work, scaling to substantially longer videos or finer temporal resolutions may require sparse, hierarchical, or coarse-to-fine lattice representations.

More broadly, our findings suggest that structured posterior modeling should be adapted to the evidence structure of each modality rather than transferred unchanged across domains. A promising direction is therefore to develop modality-specific potential and contextualization mechanisms within a shared structured-posterior framework. Such a formulation could provide a unified perspective on temporal retrieval while retaining the inductive biases required by audio, video, and potentially other sequential modalities.

7 Conclusion

We presented a structured posterior decoding framework for video moment retrieval that jointly models competing temporal spans. Building on exact segment posteriors, it contextualizes the span lattice with 2-D convolution (2DCon) and refines span potentials with posterior feedback (ITRef). The full model consistently improves over its decoding baselines across VMR benchmarks and over strong existing methods on QV-M^2, with complementary roles for the two components: 2DCon benefits queries with multiple competing moments, whereas ITRef is more effective for top-1 retrieval. Our analyses further show that the gains arise not only from global normalization, but also from how local span interactions and posterior-level retrieval beliefs are incorporated into the potential construction. The contrasting behavior on audio further suggests that span potentials should be adapted to the temporal evidence of each modality.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript and to support the retrieval and discovery of relevant prior work. We did not use generative AI tools to generate data, develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, design or implement the proposed methodology or experiments, analyze data, or interpret experimental results. We have reviewed all AI-assisted work. In particular, all AI-assisted text was reviewed and edited by the authors, and all literature suggested through AI-assisted retrieval was verified against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing benchmark datasets and pretrained representations for video and audio moment retrieval, and does not introduce a new dataset or collect new user data. As with other learning-based retrieval systems, the resulting models may inherit biases present in the underlying datasets and pretrained representations. We do not identify additional ethical concerns specific to the proposed structured retrieval methodology.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce our experiments in the appendix, including dataset statistics and evaluation protocols in Appendix A.1, and model architecture, structured-decoder configurations, training objectives, optimization settings, model-selection criteria, and computational setup in Appendix A.2. Upon acceptance, we will release the source code, configuration files, and evaluation scripts on GitHub to facilitate reproduction of the reported results.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5803–5812, 2017.

Zhuo Cao, Heming Du, Bingqing Zhang, Xin Yu, Xue Li, and Sen Wang. When one moment isn't enough: Multi-moment retrieval with cross-moment interactions. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 171705–171723, 2025.

Seungdeok Choi and Yong-Hwa Park. Multi-signal cascaded grounding for audio moment retrieval from long audio. Technical report, DCASE 2026 Challenge, 2026. URL https://dcase.community/documents/challenge2026/technical_reports/DCASE2026_Choi_62_t6.pdf.

Seungdeok Choi, Seongmin Choi, Inhan Choi, Junho Kim, Jeong-gyu Ban, and Yong-Hwa Park. Segmental posterior decoding for audio moment retrieval, 2026. URL https://arxiv.org/abs/2609.16495.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 907–916, 2026.

Cheng Huang, Yi-Lun Wu, Hong-Han Shuai, and Ching-Chun Huang. Semantic fusion augmentation and semantic boundary detection: A novel approach to multi-target video moment retrieval.

In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6783–6792, 2024.

Wang Jing, Aixin Sun, Hao Zhang, and Xiaoli Li. MS-DETR: Natural language video localization with sampling moment-moment interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1387–1400, 2023.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *International Conference on Machine Learning (ICML)*, pp. 282–289, 2001.

Jie Lei, Tamara L Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11846–11858, 2021.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffact++: Diffusion action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1644–1659, 2025.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23023–23033, 2023.

WonJun Moon, Sangeek Hyun, Subeen Lee, and Jae-Pil Heo. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984, 2026.

Hokuto Munakata, Takehiro Imamura, Taichi Nishimura, and Tatsuya Komatsu. CASTELLA: Long audio dataset with captions and temporal boundaries. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 15352–15356, 2026.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. M2D-CLAP: Exploring General-purpose Audio-Language Representations Beyond CLAP. *IEEE Access*, 13:163313–163330, 2025.

You Qin, Qilong Wu, Yicong Li, Wei Ji, Li Li, Pengcheng Cai, Lina Wei, and Roger Zimmermann. Generalized video moment retrieval. In *International Conference on Learning Representations*, pp. 49207–49223, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Sunita Sarawagi and William W Cohen. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18857–18866, 2023.

Tao Song. Onset and offset weighted loss function for sound event detection, 2024. URL `https://arxiv.org/abs/2403.13254`.

Rui Su, Wanli Ouyang, Luping Zhou, and Dong Xu. Improving action localization by progressive cross-stream cooperation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12008–12017, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXV*, pp. 396–416, 2024.

Zhengxuan Wei, Jiajin Tang, and Sibei Yang. Augmenting moment retrieval: Zero-dependency two-stage learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3401–3412, 2025.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Temporal sentence grounding in videos: A survey and future directions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10443–10465, 2023.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

Songyang Zhang, Houwen Peng, Jianlong Fu, Yijuan Lu, and Jiebo Luo. Multi-Scale 2D Temporal Adjacency Networks for Moment Localization With Natural Language. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(12):9073–9087, 2022.

Xinyang Zhou, Fanyue Wei, Lixin Duan, Angela Yao, and Wen Li. The devil is in the spurious correlations: Boosting moment retrieval with dynamic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20981–20990, 2025.

Jingwei Zhuo, Yong Cao, Jun Zhu, Bo Zhang, and Zaiqing Nie. Segment-level sequence modeling using gated recursive semi-Markov conditional random fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1413–1423, 2016.

A Appendix

A.1 Dataset and Evaluation Details

A.1.1 Dataset Statistics

Table 5 summarizes the datasets used in our experiments. QV-M^2 exhibits a pronounced distribution shift in target count: its test split contains substantially more annotated moments per query than the training and validation splits. QVHighlights and Charades-STA provide complementary conventional VMR settings, while CASTELLA is used for the cross-modal analysis.

The QVHighlights test annotations were publicly released by the dataset authors after the CodaLab evaluation server was discontinued. We evaluate on them locally with the same evaluator used for QV-M², and all checkpoint selection uses the validation split only.

Table 5: Dataset statistics. For QV-M^2, statistics are reported separately for each split because the target-count distribution differs substantially across splits.

Dataset	Split	Queries	Videos	Mom./query	Single (%)	Max mom.	Median mom. (s)
QV-M^2	Train	8,878	7,300	1.96	61.4	24	14
	Val	1,779	1,546	1.98	60.8	25	14
	Test	323	173	3.31	24.1	14	6
QVHighlights	Val	1,550	1,519	1.81	65.8	25	16
	Test	1,542	1,529	1.79	66.9	20	16
Charades-STA	Test	3,720	1,334	1.00	100.0	1	7
CASTELLA	Test	1,347	566	3.10	42.5	24	4

A.1.2 Evaluation Metrics

Average precision. For each query, average precision (AP) is computed at temporal IoU thresholds

$$\mathcal{T}_{\mathrm{AP}} = \{0.50, 0.55, \ldots, 0.95\},$$

using at most the top 10 predictions. G-mAP is the mean of query-level AP over all queries and all thresholds.

Let n_q denote the number of ground-truth moments for query q. $\mathrm{mAP@1tgt}$ and $\mathrm{mAP@2tgt}$ are the mean AP over queries satisfying $n_q = 1$ and $n_q = 2$, respectively. For queries with at least three targets, the official QV-M^2 evaluator first computes the mean AP separately for each target-count group and then averages these group means:

$$\mathrm{mAP@3{+}tgt} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{Q}_c|} \sum_{q \in \mathcal{Q}_c} \mathrm{AP}_q, \tag{13}$$

where $\mathcal{Q}_c = \{q : n_q = c\}$ and $\mathcal{C} = \{c \geq 3 : |\mathcal{Q}_c| > 0\}$. Thus, target-count groups are weighted equally rather than by their number of queries.

Multi-moment recall. For each $\tau \in \{0.30, 0.35, \ldots, 0.95\}$, the top-$k$ predictions are greedily matched one-to-one to ground-truth moments in prediction rank order. Let $m_k(q, \tau)$ be the number of matched pairs whose tIoU is at least τ. Then

$$\mathrm{mR@}k = \frac{1}{|\mathcal{T}_R|} \sum_{\tau \in \mathcal{T}_R} \frac{\sum_{q \in \mathcal{Q}_k} m_k(q, \tau)}{\sum_{q \in \mathcal{Q}_k} \min(n_q, k)}, \tag{14}$$

where $\mathcal{Q}_k$ contains queries with at least k ground-truth moments and at least k predictions.

Multi-moment localization. Using the same greedy one-to-one matching, $\mathrm{mIoU@}k$ is the mean tIoU of the matched top-k prediction–ground-truth pairs, without applying an IoU threshold. The same query eligibility criterion as $\mathrm{mR@}k$ is used.

Standard VMR recall. We additionally report R1@ τ for $\tau \in \{0.5, 0.7\}$, which measures the fraction of queries for which the top-ranked prediction overlaps at least one ground-truth moment by tIoU $\geq \tau$.

A.2 Implementation Details

A.2.1 Pretrained Features

All pretrained feature encoders are frozen throughout training. Table 6 summarizes the feature configurations used in our experiments.

Table 6: Pretrained input features used in the VMR and AMR experiments.

Setting	Modality	Encoder	Temporal rate	Dim.
VMR (main)	Video	CLIP ViT-B/32 + SlowFast	1 / 2 s	2816
	Text	CLIP	token-level	512
VMR (IV2)	Video	InternVideo2-6B	1 / 2 s	768
	Text	LLaMA	token-level	4096
AMR	Audio	M2D-CLAP	1 / 1 s	768
	Text	M2D-CLAP	token-level	768

The main VMR configuration uses the official CLIP+SlowFast and CLIP features distributed with QVHighlights/QV-M^2. Video features are extracted every 2 s, with at most 75 video tokens, and the text sequence contains at most 32 tokens. For the InternVideo2 setting, both modalities are replaced: InternVideo2-6B provides 768-dimensional video features and LLaMA provides 4096-dimensional text features. The corresponding QV-M^2 training split contains 8,876 queries rather than 8,878 because features are unavailable for two training queries; the validation and test query sets are unchanged.

For CASTELLA, we use M2D-CLAP (`m2d_clap_vit_base-80x1001p16x16p16kpBpTI-2025`) for both audio and text. Audio is represented using 1-s windows at 16 kHz.

A.2.2 QD-DETR BACKBONE

The VMR experiments use QD-DETR as the shared backbone. For each video token, two temporal endpoint features $[t/L, (t+1)/L]$ are concatenated to the pretrained video feature. Video and text features are independently projected to a common 256-dimensional space using two-layer input projections consisting of LayerNorm, dropout, a linear layer, and ReLU.

A learned global token is prepended to the joint video–text sequence, and sinusoidal positional encoding is applied to the video stream. The encoder uses two text-to-video cross-attention layers followed by two self-attention layers over the global and video tokens. After cross-modal encoding, the text tokens are removed and the structured decoder receives the resulting sequence of L query-conditioned video tokens. The pooled text representation used by the structured modules is the masked mean of the projected text tokens.

Table 7 summarizes the backbone configuration.

Table 7: QD-DETR backbone configuration used in the VMR experiments.

Configuration	**Value**
Hidden dimension	256
Input projection layers	2
Input dropout	0.5
Transformer dropout	0.1
Cross-attention layers	2
Self-attention layers	2
Attention heads	8
Feed-forward dimension	1024
DETR decoder layers	2
DETR query slots	10
Positional encoding	Sinusoidal

A.2.3 STRUCTURED DECODER CONFIGURATION

The base structured decoder uses a maximum segment duration $D_{\max} = 75$ for QV-M^2 and QVHighlights. The explicit span-mean content term and duration prior are disabled in all headline configurations. The frame-level head is a two-layer MLP,

$$256 \to 256 \to 3,$$

with GELU activation and dropout 0.1, producing onset, offset, and background scores.

For 2-D span-lattice contextualization, the video memory is projected to 32 channels. The contextualizer consists of four 2-D convolutional blocks, each composed of a convolution, GroupNorm with four groups, and ReLU, followed by a 1×1 output projection. Unless otherwise stated, we use a 3×3 kernel and contextualize spans up to $D_{\mathrm{2D}} = 64$.

Posterior-conditioned refinement projects each span to a 64-dimensional representation and performs two refinement iterations. The refinement operates on spans up to $D_{\mathrm{ref}} = 64$. Posterior weights used to construct the moment prototype are detached from the computation graph.

Initialization and masking. The final 1×1 projection of the 2-D contextualizer is zero-initialized, while its external scale is initialized to one. The refinement scale is also initialized to zero. Consequently, adding either proposed module leaves the underlying structured decoder unchanged at initialization.

Invalid or padded span potentials are masked with a large negative constant (-10^4) before dynamic programming. All potentials are masked according to the valid sequence length, and the partition function for sample b is evaluated at its valid length L_b.

Table 8: Default structured-decoder hyperparameters.

Configuration	Value
Base potential	onset + offset + FG bias
Explicit span-content term	Off
Duration prior	Off
$D_{\max}$ (QV-M^2/QVHighlights)	75
2-D channels	32
2-D convolutional layers	4
Default 2-D kernel	3×3
D_{2D}	64
Refinement dimension	64
Refinement iterations	2
D_{ref}	64
Posterior gradient in refinement	Stop-gradient

A.2.4 Effect of Refinement Iterations

Table 9: Effect of the number of posterior-conditioned refinement iterations on QV-M^2. Zero iterations correspond to 2DCon without iterative refinement. Results are mean $\pm$ standard deviation over three independently trained seeds.

Iterations	R1@0.5	R1@0.7	G-mAP	mAP by # targets		
				@1tgt	@2tgt	@3+tgt
0	53.56±1.94	42.72±2.03	34.17±0.94	48.49±1.55	42.39±0.63	24.98±0.51
1	55.83±0.18	43.96±1.55	34.93±0.96	**49.83±2.91**	42.94±1.56	25.78±1.90
2 (default)	**56.45±2.01**	**44.17±2.67**	**34.95±1.94**	46.43±3.09	**44.81±2.88**	**26.86±1.34**

We examine the effect of repeatedly feeding the structured posterior back into the span potentials. Table 9 compares zero, one, and two posterior-conditioned refinement iterations, where zero iterations corresponds to the 2DCon model without iterative refinement. A single refinement step improves all reported metrics over 2DCon alone. A second iteration adds little to G-mAP (34.93 $\rightarrow$ 34.95) but shifts the gain toward multi-target queries (@2tgt +1.9, @3+tgt +1.1) at the cost of single-target mAP.

A.2.5 Training Objectives

The structured branch is jointly optimized with the standard QD-DETR training objectives. Table 10 lists all loss coefficients used in the QV-M^2 experiments.

Table 10: Training-loss coefficients for the main VMR configuration.

Loss	Weight
L1 span regression	10
Generalized IoU	1
DETR classification	4
Saliency loss	1
Coarse auxiliary loss	0.25
Span re-ranking BCE	0.5
Segmental NLL	0.5
Minimum-IoU-risk	1.0

The Hungarian matching costs for span regression, generalized IoU, and classification are 10, 1, and 4, respectively. The DETR no-object coefficient is 0.1 and the saliency margin is 0.2. Auxiliary DETR losses are applied to intermediate decoder layers, whereas the structured losses are computed only from the final encoder representation. The coarse auxiliary loss applies the shared projection, encoder, and prediction heads to a 2× average-pooled token sequence with recomputed temporal

endpoint features, and is used only during training. The span re-ranking loss pools the encoder memory over each (detached) predicted span, projects it, and applies a binary cross-entropy to its cosine similarity with the mean-pooled projected query (Choi & Park, 2026).

A.2.6 Optimization and Model Selection

Table 11: Default optimization settings.

Configuration	**Value**
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Training epochs	200
Training batch size	32
Evaluation batch size	100
Gradient clipping	0.1
LR scheduler	StepLR
LR decay epoch	400
EMA	Off
Model selection	DETR val R1@0.7

Unless otherwise stated, all VMR models are trained for 200 epochs using AdamW with learning rate 10^{-4}, weight decay 10^{-4}, and a training batch size of 32. The global gradient norm is clipped to 0.1. Models are evaluated after every epoch.

The learning-rate scheduler is StepLR with a decay epoch of 400. Because training terminates after 200 epochs, the learning rate remains constant throughout the main QV-M^2 and QVHighlights experiments.

Checkpoints are selected according to the validation R1@0.7 of the shared DETR decoder rather than the structured channel. This criterion is fixed for all compared structured configurations.

A.2.7 Computational Setup

Training runs use one GPU per run on NVIDIA RTX 3090 or RTX A6000 GPUs. The default training and evaluation batch sizes are 32 and 100, respectively. For the memory-intensive CASTELLA configurations with a maximum structured duration of 300, the evaluation batch size is reduced to 25. No exponential moving average is used.

A.3 Full Ablation and Generalization Results

Table 12: Full three-seed results corresponding to Table 2. Values are mean ± standard deviation over three independently trained seeds (2023, 42, and 314). Base (DETR) uses conventional decoder confidence scoring. Base (SPD) uses segmental posterior decoding, 2DCon contextualizes the span lattice, ITRef applies posterior-conditioned iterative refinement, and Full combines 2DCon and ITRef.

Method	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt	mIoU@1	mIoU@2	mIoU@3	mR@1	mR@2	mR@3
(a) QV-M^2 test — CLIP+SlowFast												
Base (DETR)	49.74 ± 1.39	37.36 ± 1.82	25.48 ± 0.98	35.85 ± 0.23	34.24 ± 2.82	15.14 ± 1.73	47.79 ± 1.24	43.43 ± 1.65	38.08 ± 1.53	40.03 ± 1.14	36.11 ± 2.25	30.91 ± 2.11
Base (local-ψ)	37.36 ± 2.49	23.53 ± 0.54	16.19 ± 0.28	32.67 ± 0.58	21.40 ± 0.59	4.55 ± 0.20	38.99 ± 0.87	25.10 ± 1.53	18.19 ± 1.15	29.20 ± 0.72	19.64 ± 0.87	14.70 ± 0.52
Base (SPD)	55.93 ± 2.28	43.45 ± 2.06	32.42 ± 0.23	47.55 ± 2.21	40.85 ± 1.92	20.98 ± 0.67	53.57 ± 1.50	45.63 ± 0.45	38.34 ± 1.48	46.86 ± 1.53	40.78 ± 0.55	34.65 ± 0.62
2DCon	53.56 ± 1.94	42.72 ± 2.03	34.17 ± 0.94	48.49 ± 1.55	42.39 ± 0.63	24.98 ± 0.51	53.50 ± 1.80	46.08 ± 1.12	39.00 ± 1.42	46.34 ± 1.89	41.03 ± 0.90	35.68 ± 1.59
ITRef	56.25 ± 1.00	44.99 ± 1.09	33.69 ± 0.70	50.79 ± 2.75	42.88 ± 0.72	21.01 ± 0.72	54.76 ± 0.55	46.47 ± 0.84	38.11 ± 0.57	48.31 ± 0.93	41.66 ± 0.76	34.83 ± 0.31
Full (local-ψ)	52.53 ± 5.57	40.87 ± 3.57	28.03 ± 1.97	40.43 ± 3.42	38.68 ± 3.22	19.40 ± 1.46	51.47 ± 4.10	42.58 ± 2.64	34.42 ± 2.42	44.29 ± 3.60	37.12 ± 1.80	30.88 ± 2.01
Full (Ours)	56.45 ± 2.01	44.17 ± 2.67	34.95 ± 1.94	46.43 ± 3.09	44.81 ± 2.88	26.86 ± 1.34	54.25 ± 1.82	47.51 ± 1.91	40.95 ± 1.03	47.70 ± 2.28	43.17 ± 1.95	37.61 ± 1.14
(b) QV-M^2 test — InternVideo2-6B + LLaMA												
Base (DETR)	56.14 ± 1.70	40.35 ± 1.09	28.83 ± 0.54	45.08 ± 1.35	34.22 ± 0.88	16.13 ± 2.30	53.36 ± 1.14	46.37 ± 1.37	41.05 ± 1.00	44.31 ± 1.02	38.09 ± 1.56	34.47 ± 1.30
Base (SPD)	59.75 ± 2.15	46.34 ± 2.02	34.51 ± 0.62	56.64 ± 1.75	41.16 ± 1.30	19.51 ± 0.35	57.45 ± 1.69	47.77 ± 0.75	39.60 ± 0.83	49.42 ± 1.50	41.85 ± 0.73	35.84 ± 0.71
Full	62.02 ± 0.72	48.61 ± 1.64	38.43 ± 0.73	59.18 ± 2.86	44.60 ± 1.44	26.62 ± 0.11	59.88 ± 1.40	49.90 ± 1.49	42.52 ± 1.06	52.23 ± 1.10	44.14 ± 0.88	38.54 ± 0.85
(c) QVHighlights test												
Base (DETR)	60.29 ± 1.26	43.92 ± 0.18	38.71 ± 0.77	47.69 ± 1.23	25.42 ± 0.78	7.71 ± 0.15	55.46 ± 0.50	39.18 ± 0.52	33.98 ± 0.37	47.50 ± 0.61	31.36 ± 0.71	26.59 ± 0.37
Base (SPD)	61.28 ± 0.44	45.78 ± 0.68	42.97 ± 0.44	51.81 ± 0.67	30.47 ± 0.72	10.52 ± 0.60	57.77 ± 0.55	39.11 ± 1.04	31.92 ± 0.98	50.23 ± 0.48	33.96 ± 0.89	27.80 ± 0.46
Full	62.30 ± 0.99	47.15 ± 0.73	44.70 ± 0.51	53.91 ± 0.52	30.46 ± 1.05	13.20 ± 0.44	58.98 ± 0.34	39.92 ± 0.17	34.25 ± 0.03	51.39 ± 0.62	34.89 ± 0.20	30.29 ± 1.00
(d) Charades-STA test												
Base (DETR)	57.11 ± 0.62	35.18 ± 0.08	34.78 ± 0.26	34.78 ± 0.26	–	–	49.31 ± 0.29	–	–	40.75 ± 0.35	–	–
Base (SPD)	56.77 ± 0.53	33.68 ± 0.46	37.37 ± 0.29	37.37 ± 0.29	–	–	50.12 ± 0.10	–	–	40.65 ± 0.06	–	–
Full	57.31 ± 0.91	36.61 ± 0.97	39.50 ± 0.47	39.50 ± 0.47	–	–	50.17 ± 0.69	–	–	41.78 ± 0.77	–	–

A.4 Full Results for Span-Lattice Receptive Fields

Table 13: Full three-seed results corresponding to Table 3. Values are mean ± standard deviation over three independently trained seeds. Kernel sizes denote the 2-D convolutional kernel over the end-time × duration axes.

Neighbor interaction	Kernel	2DCon Params	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt
None	1×1	20,962	54.70 ± 0.47	43.14 ± 1.76	32.11 ± 0.36	46.85 ± 2.11	42.00 ± 2.32	20.91 ± 1.04
None (capacity-matched)	1×1	54,334	56.04 ± 2.03	45.10 ± 0.95	32.95 ± 0.12	48.70 ± 1.55	43.44 ± 1.31	19.76 ± 0.88
Duration neighbors	1×3	29,154	55.42 ± 3.28	43.34 ± 1.72	32.23 ± 0.43	43.85 ± 1.99	41.10 ± 3.57	21.86 ± 0.38
End-time neighbors	3×1	29,154	54.28 ± 1.71	42.83 ± 2.79	33.12 ± 0.96	44.85 ± 1.56	43.03 ± 1.58	24.58 ± 0.49
End + duration neighbors	3×3	53,730	56.45 ± 2.01	44.17 ± 2.67	34.95 ± 1.94	46.43 ± 3.09	44.81 ± 2.88	26.86 ± 1.34

A.5 Full Cross-Modal Results on CASTELLA

Table 14: Full three-seed results corresponding to Table 4. Values are mean ± standard deviation on the CASTELLA test set under the full-duration ($d_{\max} = 300$) setting. Kernel sizes denote the 2-D convolutional kernel over the end-time × duration axes.

Neighbor interaction	Kernel	2DCon Params	R1@0.5	R1@0.7	G-mAP	@1tgt	@2tgt	@3+tgt
Base (SPD)	–	–	52.36 ± 0.56	40.96 ± 0.16	34.11 ± 0.17	42.09 ± 0.48	35.89 ± 0.48	14.79 ± 0.61
None	1×1	20,962	51.62 ± 0.95	40.09 ± 0.26	33.98 ± 0.34	42.18 ± 0.58	36.43 ± 0.63	14.73 ± 0.15
Duration neighbors	1×3	29,154	52.86 ± 0.30	40.54 ± 0.52	34.07 ± 0.20	42.17 ± 0.67	35.42 ± 0.74	15.92 ± 0.80
End-time neighbors	3×1	29,154	51.72 ± 0.74	39.77 ± 1.01	33.47 ± 0.49	41.48 ± 0.72	35.83 ± 0.26	14.80 ± 0.19
End + duration neighbors	3×3	53,730	52.07 ± 0.51	39.25 ± 0.83	33.15 ± 0.58	40.86 ± 0.81	35.49 ± 0.61	14.93 ± 0.84