

HG-DETR: Hierarchical-Guided Detection Transformer for Video Moment Retrieval and Highlight Detection

Hongye Chen
Nanchang Hangkong University
yehongchen0316@gmail.com

Zheng Tang
Independent Researcher
tangz_cs@163.com

Tao Huang*
Xidian University
htt1001@163.com

Abstract—Video Moment Retrieval (MR) and Highlight Detection (HD) are critical tasks in video understanding, yet existing methods predominantly rely on complex task interactions, multi-modal fusion, or computationally heavy Large Language Models (LLMs), often overlooking the inherent hierarchical structure of video semantics. Videos naturally organize content into multi-level semantic hierarchies: frames form segments, segments compose scenes, and scenes interconnect to construct narratives. Conventional flat feature representations fail to capture this hierarchical organization, leading to suboptimal alignment between visual content and textual queries. To address this limitation, we propose HG-DETR, a Hierarchical-Guided Detection Transformer for Video Moment Retrieval and Highlight Detection. The core innovation lies in introducing a hierarchical video feature modeling method for the joint MR and HD task, constructing hierarchical semantic features spanning frames, segments, scenes, and stories. Furthermore, through hierarchical guidance loss and decoding mechanisms, hierarchical semantics explicitly guide and enhance the representation and decoding process of traditional global features, enabling fine-grained understanding of video semantics. Experiments on multiple benchmark datasets including QVHighlights demonstrate that HG-DETR effectively improves model performance, validating the effectiveness of our approach in MR and HD tasks.

Index Terms—video moment retrieval, highlight detection, hierarchical modeling, transformer

I. INTRODUCTION

With the widespread adoption of digital devices and the rapid development of internet technology, video has become the dominant medium for information recording and dissemination. In this context, Video Moment Retrieval (MR) [1]–[3] and Highlight Detection (HD) [4], [5] have emerged as important research directions. The MR task aims to precisely locate temporal boundaries of target segments in long videos based on natural language queries, while the HD task focuses on evaluating the saliency of video content. Due to their similarity in objectives and methodologies, recent research has begun to study them as a joint task [6]. Lei et al. constructed the first joint task dataset QVHighlights and proposed the Moment-DETR joint task framework, enabling models to simultaneously accomplish Moment Retrieval and Highlight Detection based on natural language queries, reducing computational resource waste and improving model efficiency.

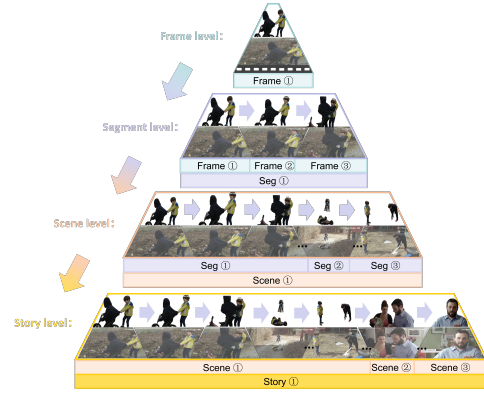


Fig. 1. Hierarchical structure in videos. From bottom to top: frames, segments, scenes, and stories. Frames combine into segments, segments organize into scenes, and scenes ultimately connect to form complete stories. The purple arrows further illustrate the action evolution relationships between different frames.

Since then, the joint task has made a series of advances: UMT [7] incorporates audio modality into multimodal modeling; UniVTG [8] emphasizes the importance of pre-training; TR-DETR [9] explores inter-task mutual benefit relationships; MS-DETR [10] further improves performance by introducing Large Language Models (LLMs). While existing methods continue to push performance boundaries, we have identified limitations in their approaches: most work focuses on designing complex interaction mechanisms, fusion strategies, task collaboration mechanisms, or introducing LLMs at high computational costs, while commonly overlooking the core and fundamental issue of video feature modeling. Current methods primarily rely on pre-trained networks such as VGG [11], C3D [12], I3D [13], or CLIP [14] combined with SlowFast [15] for video feature extraction. These methods have deficiencies in semantic modeling: traditional 3D convolutional networks are limited by receptive fields and struggle to capture long-range dependencies [16]; CLIP+SlowFast uses shallow concatenation, making it difficult to fully construct video semantics. Therefore, current video feature modeling schemes do not yet meet the requirements for feature expressiveness in joint tasks.

In fact, while videos are presented as frame sequences, the semantic organization of videos has an inherent hierar-

*Corresponding author.

chical structure [17], as shown in Figure 1: objects constitute frames, frames aggregate into segments, segments combine into scenes, and scenes connect to form stories, ultimately forming a complete video narrative. However, traditional feature modeling methods struggle to effectively capture hierarchical semantics. To address this, we propose HG-DETR: Hierarchical-Guided Detection Transformer for Video Moment Retrieval and Highlight Detection. Through hierarchical modeling methods, we construct hierarchical semantic features and utilize hierarchical features to guide the decoding process of traditional features, achieving further performance improvements. The hierarchical modeling method employs coarse-fine alternating perception and hierarchical modeling strategy. At each level, fine-grained queries align local semantics (objects, events), while coarse-grained queries maintain global context consistency. Ultimately, features at all levels are fused to form hierarchical video features. Additionally, we design hierarchical guidance loss and decoding mechanisms. Using the hierarchical guidance loss function, we align the comprehensive semantic representations of hierarchical features and traditional features through cosine similarity; through the guided decoding method, hierarchical semantics from hierarchical features guide global feature decoding, completing the overall semantic understanding of video content. The main contributions of this work are:

- We propose a hierarchical video feature modeling method for the joint MR and HD task, which employs a coarse-fine alternating perception mechanism and hierarchical modeling strategy. By alternately introducing fine-grained and coarse-grained queries at each semantic level, we construct hierarchical video features that effectively capture multi-level semantics spanning segments, scenes, stories, and complete videos.
- We design Hierarchical Guidance Loss and Decoding Mechanisms to achieve explicit semantic guidance. Through the hierarchical guidance loss function, we align hierarchical semantics with traditional features in the semantic space. Through the guided decoding method, hierarchical semantic signals effectively guide the decoding process of traditional features, strengthening both the depth and effectiveness of semantic guidance.
- We construct the HG-DETR model and conduct experiments on three datasets: QVHighlights, TVSum, and Charades-STA. Results demonstrate state-of-the-art performance across key metrics, validating the effectiveness of our approach in both MR and HD tasks.

II. RELATED WORKS

A. Video Moment Retrieval and Highlight Detection

Video moment retrieval aims to find relevant moments in videos based on natural language queries. This task typically employs two approaches: first, proposal-based methods [1], [18]–[20] that use sliding windows or proposal generation networks to create candidate windows, then select candidates with high similarity to the query; second, proposal-free methods [21], [22] that directly predict the start and end

times of relevant moments through video-text interaction via regression. Highlight detection identifies the most important parts of videos. Initially, it did not involve multimodality [23] but directly evaluated the saliency scores of video segments to obtain key segments of the entire video. Later, it evolved to perform detection based on user preferences through text queries, and subsequently introduced audio cues [7] to provide complementary information for visual features.

There exists a correlation between MR and HD. The introduction of the QVHighlights dataset combines annotation information from two independent tasks and proposes a joint task baseline called Moment-DETR [6]. Subsequently, much work has developed around this DETR [24] framework for joint task research. UMT introduces audio modality into the joint task, UniVTG emphasizes the importance of pre-training, and TR-DETR explores inter-task mutual benefit relationships. Unlike previous methods, our proposed approach focuses on hierarchical semantics in videos, achieving more detailed video understanding and more accurate predictions through hierarchical semantic guidance of traditional video features.

B. Video Feature Modeling

The core of video feature modeling lies in how to effectively and uniformly understand the spatial and temporal information within videos. Early methods relied on hand-crafted features and 2D CNNs aggregating temporal information via pooling or LSTM, though with limited long-term modeling [13], [25]. Two-stream networks fused appearance and motion via RGB and optical flow [26]. 3D CNNs (C3D, I3D) directly learned spatiotemporal features through 3D convolution [12], [13], with efficient variants like R(2+1)D decomposing 3D convolution to balance performance and efficiency [25]. In recent years, Transformer architectures based on self-attention mechanisms have gradually become mainstream, effectively capturing long-range spatiotemporal dependencies [16], [27], [28]. Currently, frontier directions in video modeling mainly include self-supervised learning and multimodal large models, aiming to reduce dependence on annotated data and fuse multi-dimensional information such as vision and language.

However, when facing multi-task joint modeling, most existing methods directly use global features extracted by pre-trained networks, including CLIP, SlowFast, C3D, I3D, and VGG, lacking modeling of the internal hierarchical semantic structure of videos, resulting in limitations in complex semantic retrieval and local event understanding tasks.

III. METHOD

We propose a hierarchical semantic guidance method for MR and HD. The overall pipeline is shown in Figure 2. We extract video static features $F^C \in \mathbb{R}^{T \times d_C}$ and dynamic features $F^S \in \mathbb{R}^{T \times d_S}$ using CLIP and SlowFast, object features $F^O \in \mathbb{R}^{T \times d_O}$ via object detection, and frame description features $F^D \in \mathbb{R}^{T \times d_D}$ from image-to-text models. For query text, we use fine-grained query features $F^{T_i} \in \mathbb{R}^{T \times d_{T_i}}$ from CLIP text encoder and coarse-grained query features $F^{T_g} \in \mathbb{R}^{T \times d_{T_g}}$ from BiGRU encoding. Based on these

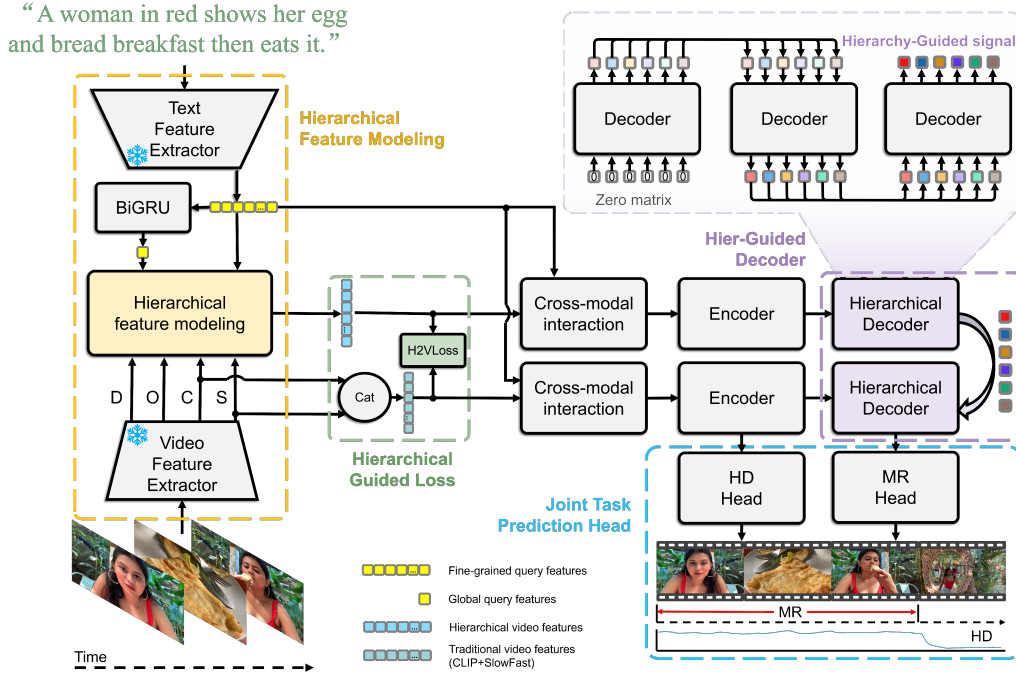


Fig. 2. Overall pipeline of HG-DETR. We propose hierarchical video feature modeling using coarse-fine alternating perception mechanism and hierarchical modeling strategy. We align traditional video features with hierarchical features through hierarchical guidance, then perform cross-modal interaction with query text for both types of features before inputting them into the Transformer encoder. Through the Hierarchical Guided Decoder, hierarchical semantic guidance signals guide traditional video features for iterative decoding. Finally, task-related predictions are made. Where D represents description features, O represents object features, C represents CLIP static features, and S represents SlowFast dynamic features.

features, we construct hierarchical video features $F^{V_h} \in \mathbb{R}^{T \times d}$ to guide traditional video features $F^{V_g} \in \mathbb{R}^{T \times d}$ (concatenation of $F^C \in \mathbb{R}^{T \times d_C}$ and $F^S \in \mathbb{R}^{T \times d_S}$). Through hierarchical guidance loss function, we bring hierarchical features and traditional features closer in semantic space. Then, after cross-modal interaction between both types of features and fine-grained text features, they are respectively input to Transformer encoders. Through the hierarchical guidance mechanism, guidance signals generated from hierarchical feature decoding are passed to traditional video features to guide their decoding process. Finally, prediction heads output predicted moments and highlight scores.

A. Hierarchical Modeling Method

To model hierarchical semantics in videos, we propose a hierarchical modeling method that performs hierarchical modeling of videos through coarse-fine alternating perception mechanism and hierarchical modeling strategy. This method constructs hierarchical semantic features spanning frames, segments, scenes, stories, and videos through semantic modeling at four levels. Specifically, first, for the input multiple video features and two granularity levels of query features, we uniformly map their dimensions to $d = 256$, obtaining dimension-mapped static features $F^{C'} \in \mathbb{R}^{T \times d}$, dynamic features $F^{S'} \in \mathbb{R}^{T \times d}$, object features $F^{O'} \in \mathbb{R}^{T \times d}$, frame description features $F^{D'} \in \mathbb{R}^{T \times d}$, fine-grained query features $F^{T'_l} \in \mathbb{R}^{T \times d}$, and coarse-grained query features $F^{T'_g} \in \mathbb{R}^{T \times d}$.

At the **frame level**, fine-grained queries interact with object features to generate preliminary representations, which are

then complemented by CLIP features via gated fusion. Coarse-grained queries perform attention modulation to obtain frame features:

$$F^{\text{frame}} = \text{MHAttn}(F^{T'_g}, \text{Gate}(\text{MHAttn}(F^{T'_l}, F^{O'}), F^{C'})), \quad (1)$$

where $F^{\text{frame}} \in \mathbb{R}^{T \times d}$ represents frame-level features.

At the **segment level**, we group consecutive frame-level features, SlowFast features, and descriptions into segment units. Following the 2-second sampling protocol commonly adopted in prior MR&HD works [6], [9], [29], video features are extracted at 2-second intervals. Based on this temporal granularity, we aggregate every 5 consecutive frames as one segment unit, spanning approximately 10 seconds. This 10-second window is designed to capture semantically coherent local actions or event evolutions. The window length is sufficient to cover the progression of a single action such as preparing ingredients or cooking, while remaining compact enough to avoid mixing multiple distinct events. For each segment unit, we perform average pooling operation along the temporal dimension, fusing static appearance and dynamic motion information across multiple frames, then further encode through Transformer encoder to generate compact segment-level features. The specific computation is as follows:

$$F^{\text{f}_{\text{seg}}} = \text{Enc}(\text{pool}(\text{bucket}(F^{\text{frame}}))), \quad (2)$$

$$F^{\text{S}_{\text{seg}}} = \text{Enc}(\text{pool}(\text{bucket}(F^{\text{S}}))), \quad (3)$$

$$F^{\text{D}_{\text{seg}}} = \text{Enc}(\text{pool}(\text{bucket}(F^{\text{D}}))), \quad (4)$$

where F^{fseg} , F^{Seg} , F^{Dseg} represent grouped frame segments, action segments, and description segments respectively, $\text{Enc}(\cdot)$ denotes Transformer Encoder encoding operation, $\text{pool}(\cdot)$ represents average pooling, and $\text{bucket}(\cdot)$ represents grouping operation. Next, we perform gated fusion between frame segment features and action segment features to achieve adaptive weighting of appearance and motion information, and use fine-grained queries for attention modulation to enhance local semantics related to query details. Subsequently, the modulated features undergo second-stage gated fusion with description segment features and receive attention modulation from coarse-grained global queries, ultimately forming hierarchically rich, query-adaptive segment-level representations. The overall computation process is as follows:

$$F^{\text{seg}} = \text{MHAttn}\left(F^{T'_g}, \text{Gate}\left(F^{\text{Dseg}}, \text{MHAttn}\left(F^{T'_l}, \text{Gate}\left(F^{\text{fseg}}, F^{\text{Seg}}\right)\right)\right)\right), \quad (5)$$

where $F^{\text{seg}} \in \mathbb{R}^{T/5 \times d}$ represents segment-level features.

At the **scene level**, we model inter-segment dependencies through Transformer encoder to generate scene-level comprehensive features. Subsequently, fine-grained queries perform attention modulation on these scene features to strengthen scene content related to specific details in the query, forming fine-grained query-aware scene-level features. The specific computation process is as follows:

$$F^{\text{scene}} = \text{MHAttn}\left(F^{T'_l}, \text{Enc}(F^{\text{seg}})\right), \quad (6)$$

where $F^{\text{scene}} \in \mathbb{R}^{T/5 \times d}$ represents scene-level features.

At the **story level**, we aggregate adjacent scenes into story units, inspired by narrative structures in film theory [30], [31] that organize content hierarchically. This captures macro-level semantic progression. Specifically, we aggregate features of the 3 scenes constituting this story unit and fuse them through Transformer encoder to generate story-level overall semantic representation. This design ensures macro-narrative completeness while avoiding semantic dilution caused by spanning multiple narrative stages, making story-level representation more focused on high-level semantic modeling and query intention matching. Finally, coarse-grained global queries perform attention modulation on story features to capture macro-narrative structure matching query intention, forming the final story-level features. The computation process is as follows:

$$F^{\text{story}} = \text{MHAttn}\left(F^{T'_g}, \text{Enc}(\text{pool}(\text{bucket}(F^{\text{scene}})))\right), \quad (7)$$

where $F^{\text{story}} \in \mathbb{R}^{T/5/3 \times d}$ represents story-level features.

The gated fusion mechanism in the hierarchical modeling method is specifically as follows: For input features x and y , we first concatenate them along the temporal T dimension $[x; y]$, then generate gate weights and feature residuals through two parallel feedforward networks respectively. The detailed computation process is as follows:

$$\alpha = \sigma(W_2 \cdot \text{ReLU}(W_1[x; y])), \quad (8)$$

$$\Delta = W_4 \cdot \text{ReLU}(W_3[x; y]), \quad (9)$$

where $\alpha \in \mathbb{R}^{T \times 1}$ represents gate weights for controlling the degree of feature fusion, $\Delta \in \mathbb{R}^{T \times d}$ represents feature adjustment learned from $[x; y]$, σ is the Sigmoid activation function constraining gate weights to the $[0, 1]$ interval, $\text{ReLU}(\cdot)$ is the nonlinear activation function enhancing feature expressiveness, and $W_i, i = 1, 2, 3, 4$ are learnable weights. The final output is computed through residual form, specifically calculated as follows:

$$\text{output} = x + \alpha \otimes \Delta, \quad (10)$$

where $\otimes$ denotes element-wise multiplication of scalar to vector, and $\text{output} \in \mathbb{R}^{T \times d}$ is the final fusion output result.

In the final stage of modeling, semantic information at each level is backpropagated to frame level. Segment, scene, and story representations are respectively upsampled or broadcast to the original frame sequence length, concatenated with frame-level features, and fused through multi-layer perceptron to output the final representation of **hierarchical features** $F^{V_h} \in \mathbb{R}^{T \times d}$. This enables each frame feature to simultaneously contain local details, segment actions, coherent scene, and story global semantics, achieving hierarchical video feature modeling. Meanwhile, we still preserve **traditional features** $F^{V_g} \in \mathbb{R}^{T \times d}$ following the previous work's approach of concatenating CLIP features with SlowFast features.

B. Hierarchical Guidance Loss and Decoding Mechanism

Unlike traditional approaches that directly fuse two types of features, we propose utilizing hierarchical semantics contained in hierarchical features to guide the decoding process of traditional features. This design is primarily based on the following considerations: direct fusion indiscriminately merges features at different abstraction levels, which, while preserving all information, also introduces noise and interference. Especially for MR tasks that are sensitive to temporal boundaries, local details are easily overwhelmed by global context. Additionally, the model needs to learn how to parse effective representations for respective sub-tasks from mixed features, which increases learning difficulty and may lead to optimization conflicts. Subsequent ablation experiments also provide corresponding experimental evidence of the performance impact of direct fusion. Based on these ideas, we design hierarchical guidance loss and decoding mechanism. This mechanism consists of two synergistic components: Guidance Loss and Guided Decoding Method, forming a path from feature space constraint to decoding process guidance.

1) *Guidance Loss*: The core objective of Guidance Loss is to help hierarchical features guide traditional features. We directly perform average pooling on both types of features along the temporal dimension T to obtain their respective global feature vectors $\bar{F}^{V_h}$ and $\bar{F}^{V_g}$, which aggregate the semantics of the entire video. Then we implement semantic guidance of hierarchical features on video features through cosine embedding loss, forcing the two types of features to maintain consistency in high-level semantics and providing a more discriminative feature foundation for subsequent MR and HD.

Specifically, we use cosine embedding loss function, aiming to maximize the cosine similarity between the two types of features. This loss function is defined as:

$$\mathcal{L}_{h2v} = 1 - \frac{1}{B} \sum_{i=1}^B \frac{\bar{F}_i^{V_h} \cdot \bar{F}_i^{V_g}}{\|\bar{F}_i^{V_h}\| \|\bar{F}_i^{V_g}\|}, \quad (11)$$

where $\bar{F}^{V_h}$ and $\bar{F}^{V_g}$ are the global average vectors of hierarchical features and traditional features respectively, B is the batch size, and i represents the i -th batch. After being guided by the guidance loss, hierarchical features and traditional features respectively undergo bidirectional attention cross-modal interaction with text [9], then enter the encoder and obtain encoded hierarchical features $F^{V_h'}$ and encoded traditional features $F^{V_g'}$.

2) *Guided Decoding Method*: To fully decode the structured semantics in hierarchical features and enable them to effectively guide traditional feature decoding, inspired by [41], we improve the DETR decoder by designing a Hierarchical Decoder. This fully parses semantic information at different hierarchical levels while improving the decoding quality of traditional video features and the accuracy of hierarchical information guidance.

In the guided decoding method, we decode hierarchical features and traditional features through different Hierarchical Decoders. During the hierarchical feature decoding process, consistent with previous methods, its moment query still uses learnable vectors initialized with zero matrices. Through hierarchical decoding by the Hierarchical Decoder, we obtain hidden states rich in local object action semantics and temporal structure information after hierarchical feature decoding. The detailed calculation is as follows:

$$tgt_h[i, j, k] = 0, \quad (12)$$

$$hs_h = \text{HiDecoder}_h(tgt_h), \quad (13)$$

where $tgt_h \in \mathbb{R}^{2 \times n \times d}$ is the initial moment query initialized as a zero tensor [6], with $i \in \{1, 2\}$, $j \in \{1, \dots, n\}$, $k \in \{1, \dots, d\}$. $hs_h \in \mathbb{R}^{2 \times n \times d}$ is the hidden layer state decoded from hierarchical features, n is the number of queries, and $\text{HiDecoder}_h(\cdot)$ is the Hierarchical Decoder corresponding to hierarchical features. $hs_h \in \mathbb{R}^{2 \times n \times d}$ is the hidden layer state decoded from hierarchical features, n is the number of queries, and $\text{HiDecoder}_h(\cdot)$ is the Hierarchical Decoder corresponding to hierarchical features.

$$tgt_g = hs_h, \quad (14)$$

$$hs_g, ref_g = \text{HiDecoder}_g(tgt_g), \quad (15)$$

where $tgt_g \in \mathbb{R}^{2 \times n \times d}$ is the initial moment query for traditional feature decoding, $\text{HiDecoder}_g(\cdot)$ is the Hierarchical Decoder corresponding to traditional features. The finally decoded hidden layer state $hs_g \in \mathbb{R}^{2 \times n \times d}$ and reference point coordinates $ref_g \in \mathbb{R}^{2 \times n \times 2}$ are input to the prediction head to complete final prediction.

C. Prediction Heads and Training Loss

The prediction heads for MR and HD follow the design in [35]. The overall loss includes MR loss $\mathcal{L}_{MR}$ and HL loss $\mathcal{L}_{HL}$ [29], local loss $\mathcal{L}_{local}$ and global loss $\mathcal{L}_{global}$ for cross-modal alignment [9], as well as the Guidance Loss $\mathcal{L}_{h2v}$ proposed in this paper. The expression for overall loss $\mathcal{L}_{total}$ is:

$$\mathcal{L}_{total} = \mathcal{L}_{MR} + \mathcal{L}_{HL} + \lambda_{local} \mathcal{L}_{local} + \lambda_{global} \mathcal{L}_{global} + \lambda_{h2v} \mathcal{L}_{h2v}, \quad (16)$$

where λ_{local} , λ_{global} , λ_{h2v} are the weights for corresponding losses. These losses jointly constrain the model during training, ultimately improving its performance.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

We evaluate the performance of our proposed model on three datasets: the joint task dataset QVHighlights, the MR dataset Charades-STA, and the HD dataset TVSum.

- **QVHighlights**: This dataset supports both MR and HD tasks, containing over 10,000 YouTube videos with annotations including manually written natural language queries, query-related moments, and saliency scores for video segments.
- **Charades-STA**: A moment retrieval dataset built upon Charades, containing 16,128 query-moment pairs for localizing moments in untrimmed videos.
- **TVSum**: A video summarization dataset with 50 YouTube videos across 10 categories, providing frame-level importance scores.

The evaluation metrics for experiments are the same as in previous work [6]–[10], [29], [35].

B. Implementation Details

The parameter settings for each dataset in the experiments are as follows: For the QVHighlights dataset, we set the training epochs to 200, batch size to 32, learning rate to 1e-4, random seed to 3047, and the weights for local loss, global loss, and guidance loss to 0.3, 0.5, and 0.5 respectively. For the Charades-STA dataset, training epochs are set to 100, batch size to 32, learning rate to 1e-4, random seed to 42, and the weights for local loss, global loss, and guidance loss to 0.3, 0.5, and 0.5 respectively. For the TVSum dataset, we initially adopted the same base parameter configuration as Charades-STA. However, its small scale prevented the use of large learning rates with our method, so we reduced the learning rate appropriately and removed the guidance loss for subsequent experiments. Additionally, our model was implemented using the PyTorch framework, and all experiments were conducted on Nvidia RTX 3090 GPU and Intel Core i9-14900K CPU.

C. Experimental Results

Joint Task Experimental Analysis: We first evaluate HG-DETR on the QVHighlights test set and compare its performance with existing mainstream methods, as shown in

TABLE I
PERFORMANCE COMPARISON FOR VIDEO MOMENT RETRIEVAL AND VIDEO HIGHLIGHT TASK ON THE QVHIGHLIGHTS TEST SET. ‘†’ DENOTES TRAINING WITH AUDIO MODALITY. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Method	Moment Retrieval					Highlight Detection	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
Moment-DETR [6]	52.89	33.02	54.82	29.40	30.73	35.69	55.60
UMT [†] [7]	56.23	41.18	53.84	37.01	36.12	38.18	59.99
UniVTG [8]	58.86	40.86	57.60	35.59	35.47	38.20	60.96
LSJT [32]	60.51	41.50	56.33	36.70	36.66	39.13	61.22
MH-DETR [33]	60.05	42.48	60.75	38.13	38.38	38.22	60.51
EaTR [34]	61.36	45.79	61.86	41.91	41.74	37.15	58.65
QD-DETR [29]	62.40	44.98	62.52	39.88	39.86	38.94	62.40
CG-DETR [35]	<u>65.43</u>	48.38	64.51	42.77	42.86	40.33	66.21
TR-DETR [9]	64.66	48.96	63.98	43.73	42.62	39.91	63.42
Ms-DETR [10]	64.72	48.77	66.41	44.91	44.89	<u>40.45</u>	65.95
Ours	66.60	49.35	<u>66.07</u>	<u>44.02</u>	<u>43.51</u>	41.15	<u>66.02</u>

TABLE II
COMPARISON WITH REPRESENTATIVE HIGHLIGHT DETECTION METHODS ON THE TVSUM. ‘†’ DENOTES TRAINING WITH AUDIO MODALITY. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Model	TVSum										
	VT	VU	GA	MS	PK	PR	FM	BK	BT	DS	Avg.
SG [36]	41.1	46.2	46.3	47.7	44.8	46.1	45.2	40.6	47.1	45.5	45.1
Trailer [37]	61.3	54.6	65.7	60.8	59.1	70.1	58.2	64.7	65.6	68.1	62.8
SL-Module [38]	86.5	68.7	74.9	86.2	79.0	63.2	58.9	72.6	78.9	64.0	73.3
TCG [†] [39]	85.0	71.4	81.9	78.6	80.2	75.5	71.6	77.3	78.6	68.1	76.8
UniVTG [8]	83.9	85.1	89.0	80.1	84.6	87.0	70.9	91.7	73.5	69.3	81.0
UMT [†] [7]	87.5	81.5	88.2	78.8	81.4	87.0	76.0	86.9	84.4	79.6	83.1
CG-DETR [35]	86.9	88.8	94.8	87.7	86.7	89.6	74.8	<u>93.3</u>	<u>89.2</u>	<u>75.9</u>	86.8
QD-DETR [29]	88.2	87.4	85.6	85.0	85.8	86.9	76.4	91.3	<u>89.2</u>	73.7	85.0
TR-DETR [9]	89.3	93.0	94.3	85.1	88.0	88.6	80.4	91.3	89.5	81.6	88.1
MS-DETR [10]	89.84	93.33	<u>94.79</u>	<u>88.73</u>	<u>88.52</u>	89.02	80.5	93.96	88.73	76.88	88.43
Ours	87.54	93.33	93.88	90.00	89.98	<u>89.47</u>	77.06	90.95	88.70	82.51	<u>88.34</u>

TABLE III
PERFORMANCE COMPARISON FOR VIDEO MOMENT RETRIEVAL TASK ON THE CHARADES-STA TEST SET. ‘†’ DENOTES TRAINING WITH AUDIO MODALITY. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Model	Charades-STA		
	R@1		
	IoU=0.3	IoU=0.5	IoU=0.7
UMT [†] [7]	–	48.31	29.25
SimVTP [40]	–	44.70	26.30
QD-DETR [29]	–	57.31	32.55
UniVGT [8]	70.81	58.01	35.65
CG-DETR [35]	70.40	58.40	<u>36.30</u>
TR-DETR [9]	–	57.61	33.52
Ms-DETR [10]	<u>71.34</u>	59.62	36.48
Ours	71.85	57.81	32.86

Table I. In MR, HG-DETR improves R1@0.5 by 1.17% and R1@0.7 by 0.39%, achieving state-of-the-art performance. Additionally, it achieves second-best results in mAP@0.5, mAP@0.75, and mAPAvg. In HD, HG-DETR improves mAP by 0.7%, achieving state-of-the-art performance, and achieves second-best in HIT@1.

Video Moment Retrieval Task Experimental Analysis: We evaluate HG-DETR on the Charades-STA test set and compare its performance with existing mainstream methods,

as shown in Table III. R@1 at IoU=0.3 shows a relative improvement of 0.51%, achieving state-of-the-art performance.

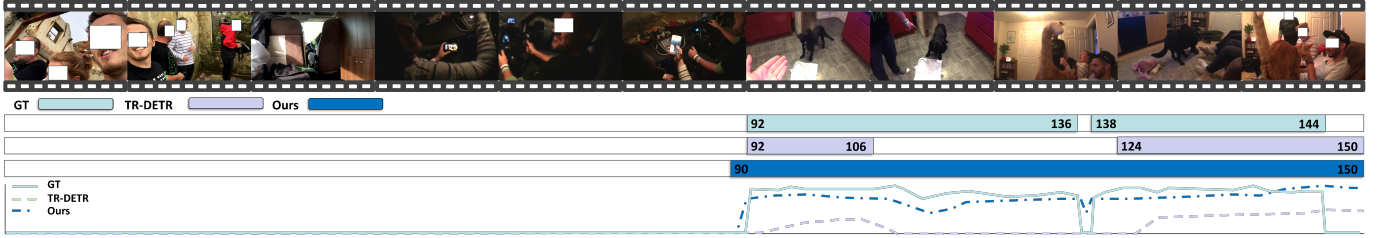
Highlight Detection Task Experimental Analysis: We evaluate HG-DETR on the TVSum test set and compare its performance with existing mainstream methods, as shown in Table II. Our method achieves state-of-the-art performance in categories VU, MS, PK, and DS. Additionally, it achieves second-best in PR and average value.

The above experimental results all demonstrate the effectiveness of our method in MR and HD tasks, further proving the effective assistance of hierarchical features for the model’s understanding of video hierarchical semantics.

D. Qualitative Results

In Figure 3, we visualize the qualitative results of HG-DETR on relevant videos from the QVHighlights dataset. Compared to our selected baseline TR-DETR, HG-DETR demonstrates more accurate moment localization and highlight prediction. In the MR task, its prediction of corresponding segment ranges is more accurate. Additionally, in the HD task, its predicted highlight scores better match the annotations, both in terms of highlight score values and highlight ranges being more precise. This indicates that HG-DETR has a clearer understanding of video segments and scenes, while TR-

(a)Query: A group of friends are playing with a black dog.



(b)Query: The woman is preparing a bowl of cereal, honey and bananas.

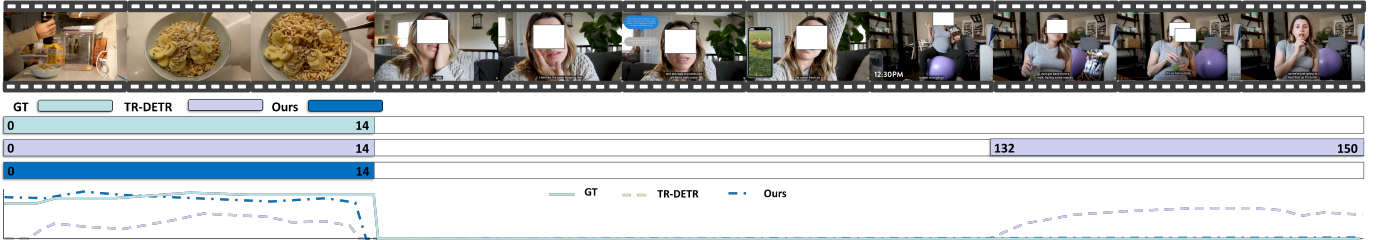


Fig. 3. Comparison of prediction results for relevant video segments between HG-DETR and TR-DETR on the QVHighlights dataset. Light blue, light purple, and deep blue represent Ground Truth, our baseline TR-DETR, and HG-DETR experimental result visualizations, respectively.

TABLE IV
PERFORMANCE COMPARISON FOR VIDEO MOMENT RETRIEVAL AND VIDEO HIGHLIGHT TASK ON THE QVHIGHLIGHTS VAL SET. ‘*’ DENOTES TRAINING WITH FEATURE FUSION. THE BEST AND SECOND-BEST RESULTS ARE SHOWN IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Method	Moment Retrieval					Highlight Detection	
	R1		mAP			$\geq$ Very Good	
	@0.5	@0.7	@0.5	@0.75	Avg.	mAP	HIT@1
HG-DETR*	68.06	49.68	66.58	42.61	43.06	41.58	67.23
HG-DETR	69.23	50.71	67.85	45.06	44.57	42.31	69.42

TABLE V
THE ABLATION RESULTS OF THE COMPONENTS OF OUR PROPOSED METHOD. ‘H2V’ REPRESENTS GUIDANCE LOSS. ‘HiD’ REPRESENTS HIERARCHICAL DECODER.

H2V	HiD	MR		HD	
		R1	Avg.	mAP	HIT@1
		@0.5	mAP	mAP	HIT@1
✓		68.32	43.61	41.95	68.06
	✓	68.58	43.82	41.19	68.19
✓	✓	69.23	44.57	42.31	69.42

DETR, lacking hierarchical feature guidance, shows deviations in relevant segment localization in Figure 3(b). It fails to distinguish between the segment of “preparing a bowl of cereal, honey and bananas” and the final segment where the lady eats and drinks water, leading to also including the final segment as a predicted segment.

E. Ablation Study

We also conducted ablation experiments on the effectiveness of relevant components in the model, as shown in Table V, demonstrating the effectiveness of each component in HG-DETR. We conducted relevant ablation experiments on the considerations mentioned in Section III-B, validating the effectiveness of using hierarchical features to guide traditional features rather than fusion. Specifically, after completing the

alignment of hierarchical features and traditional features through guidance loss, we concatenated the two types of features in the feature dimension and mapped them to 256 dimensions, and input this simple fused feature into subsequent encoders and decoders. The final metrics obtained are shown in Table IV. The performance difference between fusion results and hierarchical guidance results also validates our considerations: direct fusion of two types of features cannot fully help the model understand hierarchical semantics in videos, but instead leads to a sharp decline in model performance across most metrics.

V. CONCLUSION

Addressing the common problem of overlooking hierarchical semantic structure in video content, this paper proposes HG-DETR, a Detection Transformer framework based on hierarchical video feature modeling and guided decoding. By explicitly constructing multi-level semantic representations from frame to story level, HG-DETR can systematically capture the semantic organizational relationships within videos, providing structured semantic support for the traditional video feature decoding process. Through Hierarchical Guidance Loss and Decoding Mechanism, hierarchical semantics effectively guide the decoding of traditional features, thereby enhancing the model’s understanding of local details and global structure. Ex-

perimental results on QVHighlights, TVSum, and CharadesSTA validate the effectiveness of our proposed approach in both joint and independent MR and HD tasks, demonstrating that hierarchical modeling provides strong support for complex video semantic understanding.

REFERENCES

- [1] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5267–5275.
- [2] Z. Liu, J. Li, H. Xie, P. Li, J. Ge, S.-A. Liu, and G. Jin, "Towards balanced alignment: Modal-enhanced semantic modeling for video moment retrieval," in *Proceedings of the AAAI conference on artificial intelligence*, 2024, pp. 3855–3863.
- [3] Y. Xu, Y. Sun, B. Zhai, M. Li, W. Liang, Y. Li, and S. Du, "Zero-shot video moment retrieval via off-the-shelf multimodal large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, pp. 8978–8986.
- [4] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5179–5187.
- [5] D. Han, S. Seo, E. Park, S.-U. Nam, and N. Kwak, "Unleash the potential of clip for video highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8275–8279.
- [6] J. Lei, T. L. Berg, and M. Bansal, "Detecting moments and highlights in videos via natural language queries," *Advances in Neural Information Processing Systems*, pp. 11 846–11 858, 2021.
- [7] Y. Liu, S. Li, Y. Wu, C.-W. Chen, Y. Shan, and X. Qie, "Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3042–3051.
- [8] K. Q. Lin, P. Zhang, J. Chen, S. Pramanick, D. Gao, A. J. Wang, R. Yan, and M. Z. Shou, "Univgt: Towards unified video-language temporal grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2794–2804.
- [9] H. Sun, M. Zhou, W. Chen, and W. Xie, "Tr-detr: Task-reciprocal transformer for joint moment retrieval and highlight detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, pp. 4998–5007.
- [10] H. Ma, G. Wang, F. Yu, Q. Jia, and S. Ding, "Ms-detr: Towards effective video moment retrieval and highlight detection by joint motion-semantic learning," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 4514–4523.
- [11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [12] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.
- [13] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.
- [15] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.
- [16] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.
- [17] J. Xiao, A. Yao, Z. Liu, Y. Li, W. Ji, and T.-S. Chua, "Video as conditional graph hierarchy for multi-granular question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 2804–2812.
- [18] R. Ge, J. Gao, K. Chen, and R. Nevatia, "Mac: Mining activity concepts for language-based temporal localization," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 245–253.
- [19] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 843–851.
- [20] H. Xu, K. He, B. A. Plummer, L. Sigal, S. Sclaroff, and K. Saenko, "Multilevel language and vision integration for text-to-clip retrieval," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 9062–9069.
- [21] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 810–10 819.
- [22] K. Li, D. Guo, and M. Wang, "Proposal-free video grounding with contextual pyramid network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1902–1910.
- [23] T. Badamdorj, M. Rochan, Y. Wang, and L. Cheng, "Contrastive learning for unsupervised video highlight detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 042–14 052.
- [24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, 2020, pp. 213–229.
- [25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.
- [26] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.
- [27] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, 2021, p. 4.
- [28] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y.-G. Jiang, L. Zhou, and L. Yuan, "Bert: Bert pretraining of video transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 733–14 743.
- [29] W. Moon, S. Hyun, S. Park, D. Park, and J.-P. Heo, "Query-dependent video representation for moment retrieval and highlight detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 23 023–23 033.
- [30] S. Field, *Screenplay: The foundations of screenwriting*. Delta, 2005.
- [31] J. E. Cutting, "Narrative theory and the dynamics of popular movies," *Psychonomic bulletin & review*, pp. 1713–1743, 2016.
- [32] R. Wang, J. Feng, F. Zhang, X. Luo, and Y. Luo, "Modality-aware heterogeneous graph for joint video moment retrieval and highlight detection," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 8896–8911, 2024.
- [33] Y. Xu, Y. Sun, B. Zhai, Y. Jia, and S. Du, "Mh-detr: Video moment and highlight detection with cross-modal transformer," in *2024 International Joint Conference on Neural Networks*, 2024, pp. 1–8.
- [34] J. Jang, J. Park, J. Kim, H. Kwon, and K. Sohn, "Knowing where to focus: Event-aware transformer for video grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 846–13 856.
- [35] W. Moon, S. Hyun, S. B. Lee, and J.-P. Heo, "Correlation-guided query-dependency calibration in video representation learning for temporal grounding," *CoRR*, 2023.
- [36] Y. Yuan, L. Ma, and W. Zhu, "Sentence specified dynamic video thumbnail generation," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 2332–2340.
- [37] L. Wang, D. Liu, R. Puri, and D. N. Metaxas, "Learning trailer moments in full-length movies with co-contrastive attention," in *European Conference on Computer Vision*, 2020, pp. 300–316.
- [38] M. Xu, H. Wang, B. Ni, R. Zhu, Z. Sun, and C. Wang, "Cross-category video highlight detection via set-based learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7970–7979.
- [39] Q. Ye, X. Shen, Y. Gao, Z. Wang, Q. Bi, P. Li, and G. Yang, "Temporal cue guided video highlight detection with low-rank audio-visual fusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7950–7959.
- [40] Y. Ma, T. Yang, Y. Shan, and X. Li, "Simvtp: Simple video text pre-training with masked autoencoders," *arXiv preprint arXiv:2212.03490*, 2022.
- [41] P. Zhao, Z. He, F. Zhang, S. Lin, and F. Zhou, "Ld-detr: Loop decoder detection transformer for video moment retrieval and highlight detection," *arXiv preprint arXiv:2501.10787*, 2025.