Where Does It Exist from the Low-Altitude: Spatial Aerial Video Grounding

Yang Zhan[1], Yuan Yuan*[1]

[1] iOPEN, Northwestern Polytechnical University

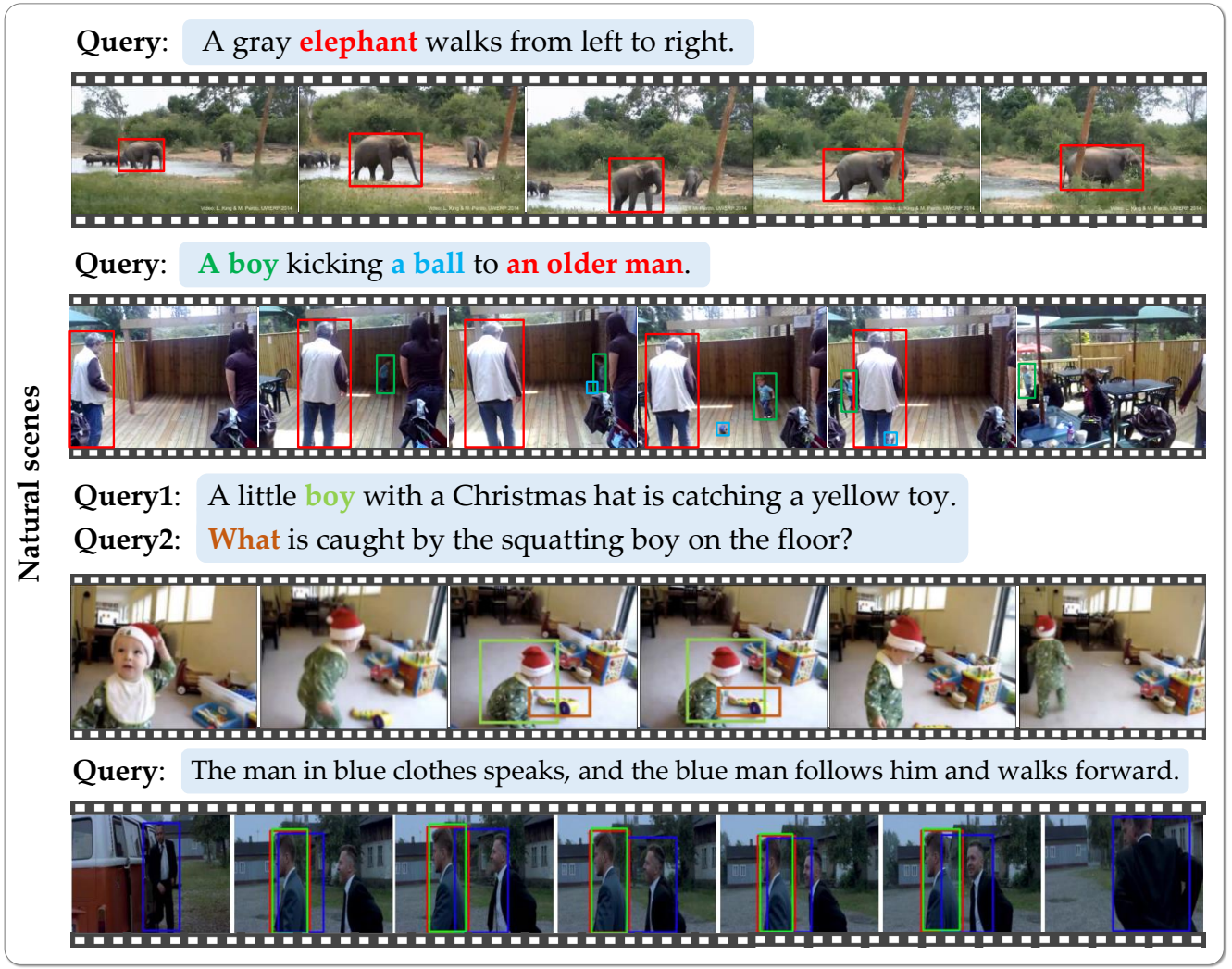
* Corresponding Author

(Contact: zhanyangnwpu@gmail.com)

Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

Grounding objects with natural language in visual contexts is a fundamental and important task in multi-modal understanding. Spatial aerial video grounding (SAVG) emerges as a groundbreaking task.



- (a) The spatial video grounding task has predominantly focused on the ego-centric fixed front perspective and simple scene. They only provide **a very limited view and environment**. This means that existing methods can only perform target localization in simple scenes on the ground. This **overlooks another important application scenario: moving aerial platforms in the sky.**

- (b) As the low-altitude economy takes off, many tasks currently need to be performed in the sky, such as UAV-based goods delivery, traffic/security patrol, and scenery tours. We can **use UAVs to localize specific objects in various scenarios from the sky view** and obtain a much more holistic grounding.

Motivation

Spatial aerial video grounding (SAVG) grounds the referred object's spatial tube in the complex aerial scene by a natural language query.

SAVG in moving aerial platforms is quite different, presenting unique technical challenges.

UAV-SVG

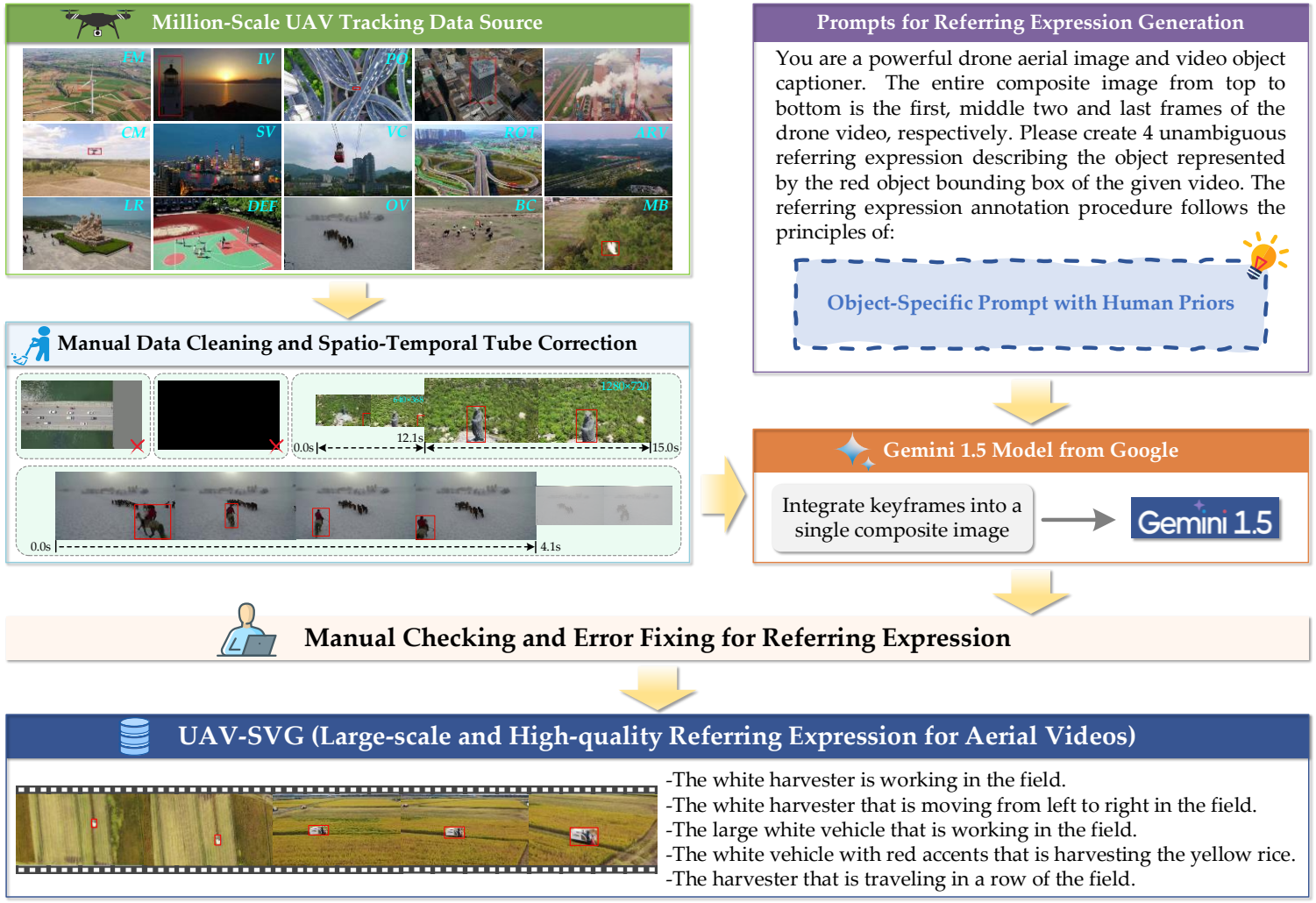
Query: The man riding the electric bike next to the flower bed form the upper left corner of the roundabout to the lower left.

Query: The green bus that is heading to the upper right, from the lower left corner of the Bell Tower roundabout.

Dataset: UAV-SVG

NEURAL INFORMATION PROCESSING SYSTEMS

Overview of data construction pipeline: Step 1) manual data cleaning and spatio-temporal tube correction, Step 2) referring expression generation, and Step 3) manual checking and error fixing.



Dataset: UAV-SVG

NEURAL INFORMATION PROCESSING SYSTEMS

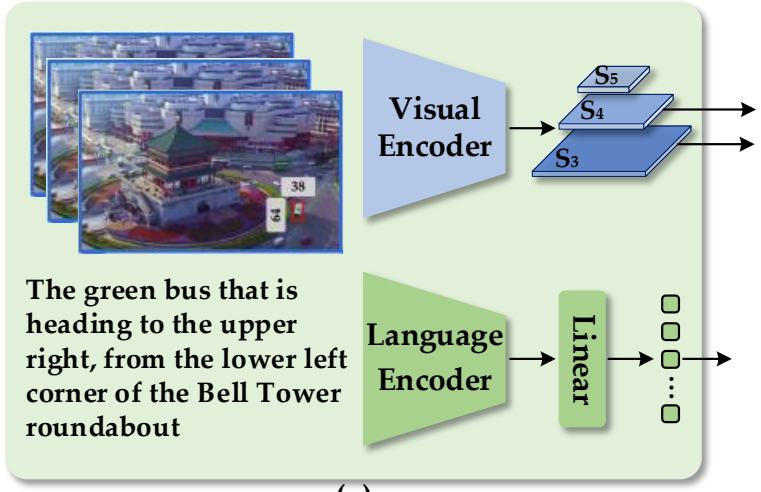
Statistic comparison of spatial video grounding datasets: UAV-SVG contains over 2 million frames, 17,820 video–query pairs, and 216 highly diverse object categories.

Dataset	Videos Num.	Frame Num.	Total Duration	Object Classes	Motion Classes	Language Num.	Vocab	Exp. Length	Train/Val/Test Partition	Target or Scene	Shot & View
VID-sentence [6]$_{ACL'19}$	7,654	59K	11.1 h	30	✗	7,654	1,823	13.18	86%/7%/7%	Animal & Vehicle	Fixed / Handheld & Front
VidSTG [7]$_{CVPR'20}$	6,924	7.1M	69.1 h	79	✗	99,943	1,881	10.12	80%/10%/10%	Human & Animal	Fixed / Handheld & Front
HC-STVG [8]$_{TCSVT'21}$	5,660	3M	31.4 h	1	✗	5,660	2,289	17.27	80%/0%/20%	Human	Movie Clips
UAV-SVG	**3,564**	**2M**	**18.7 h**	**216**	**73**	**17,820**	**3,243**	**16.39**	**79%/5%/16%**	**Wild**	**Moving Aerial & Bird's-Eye**

'Exp.' indicates expression.

Method

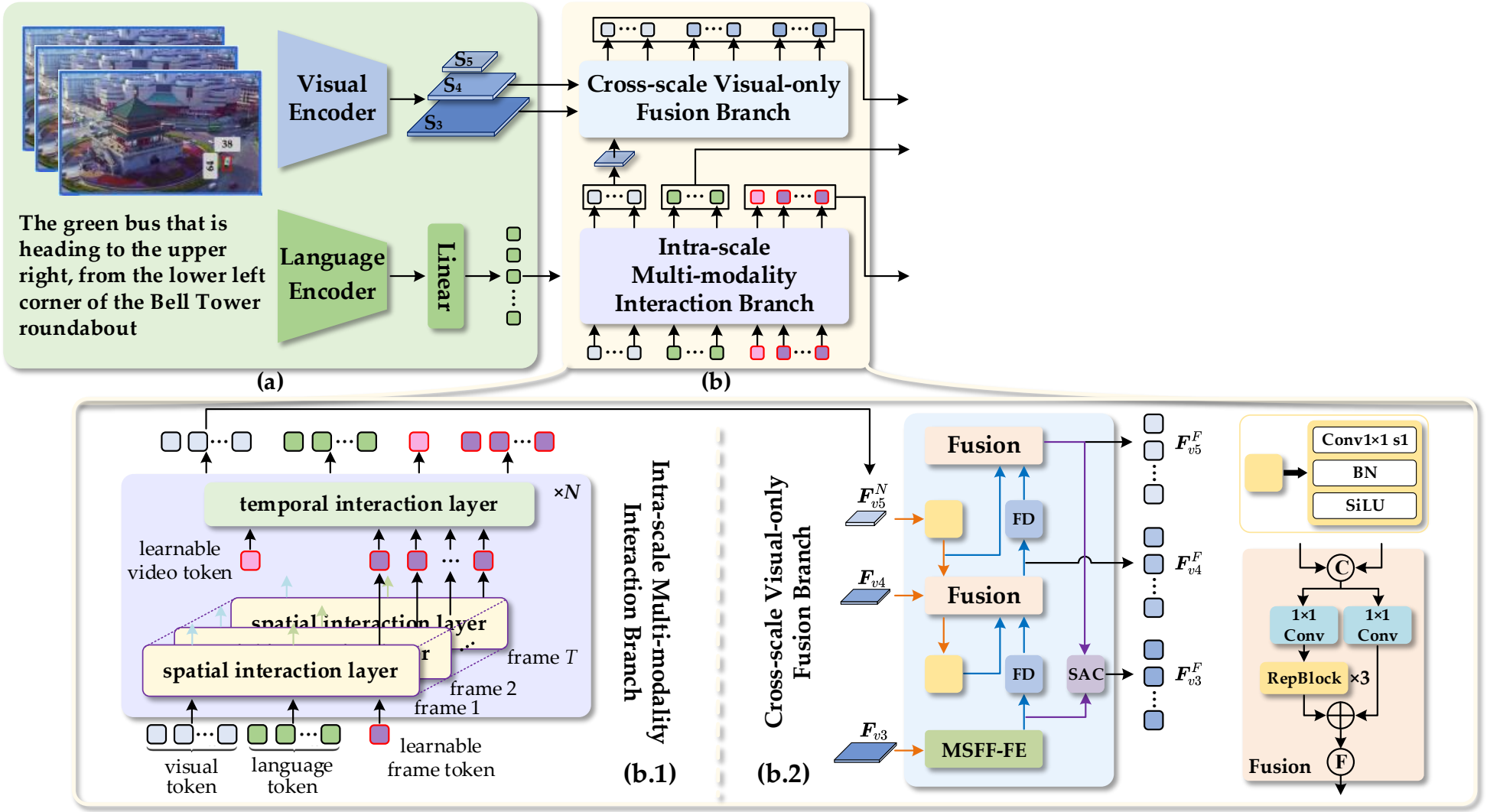
Overview of SAVG-DETR: First, the Aerial Video-Text Feature Extractor extracts aerial visual features for each frame and language features.



(a)

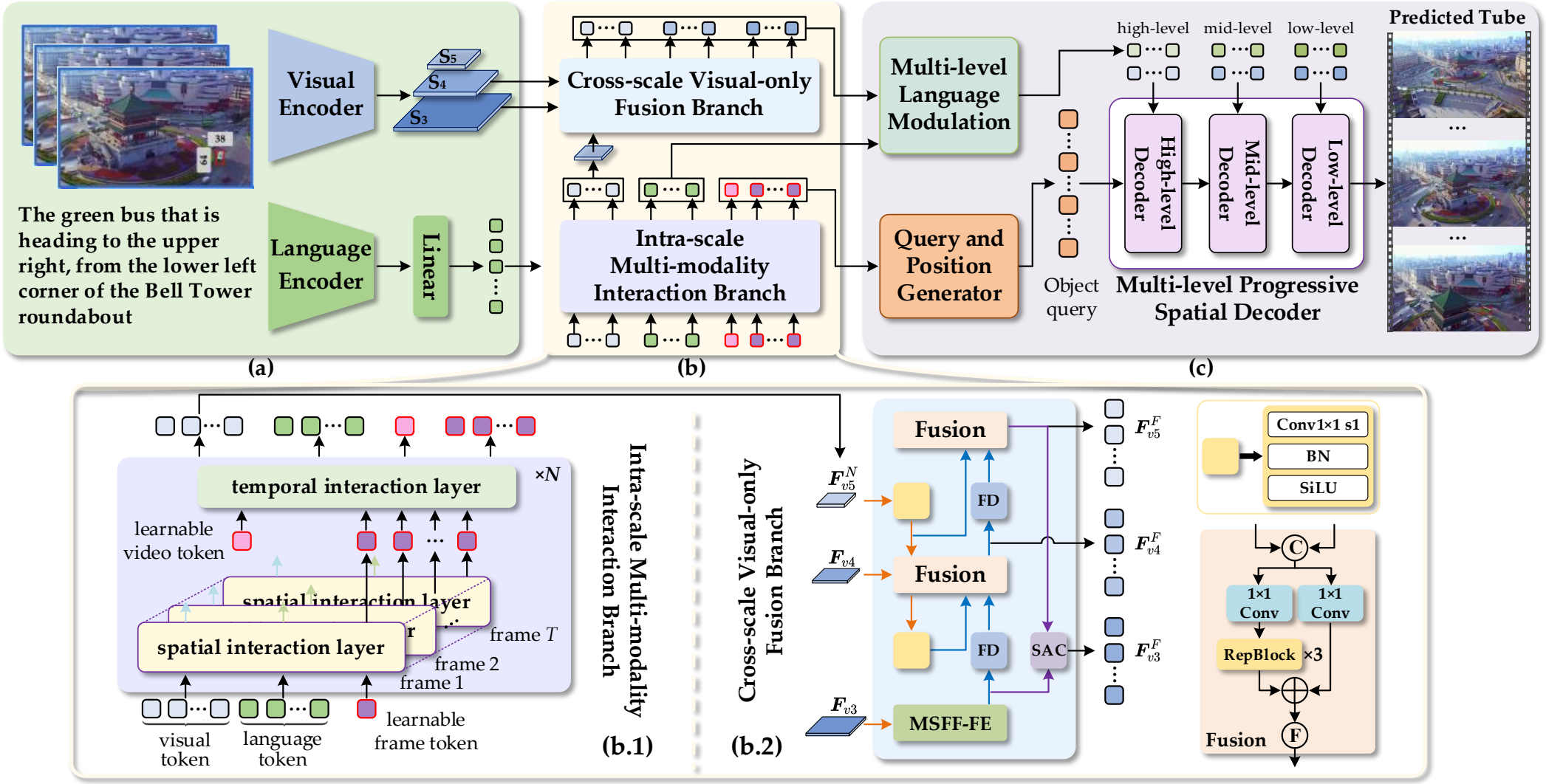
Method

Overview of SAVG-DETR: Then, the Multi-Modality Multi-Scale Spatio-Temporal Encoder can capture conceptual entities on high-level features and integrate more object details from low-level features.



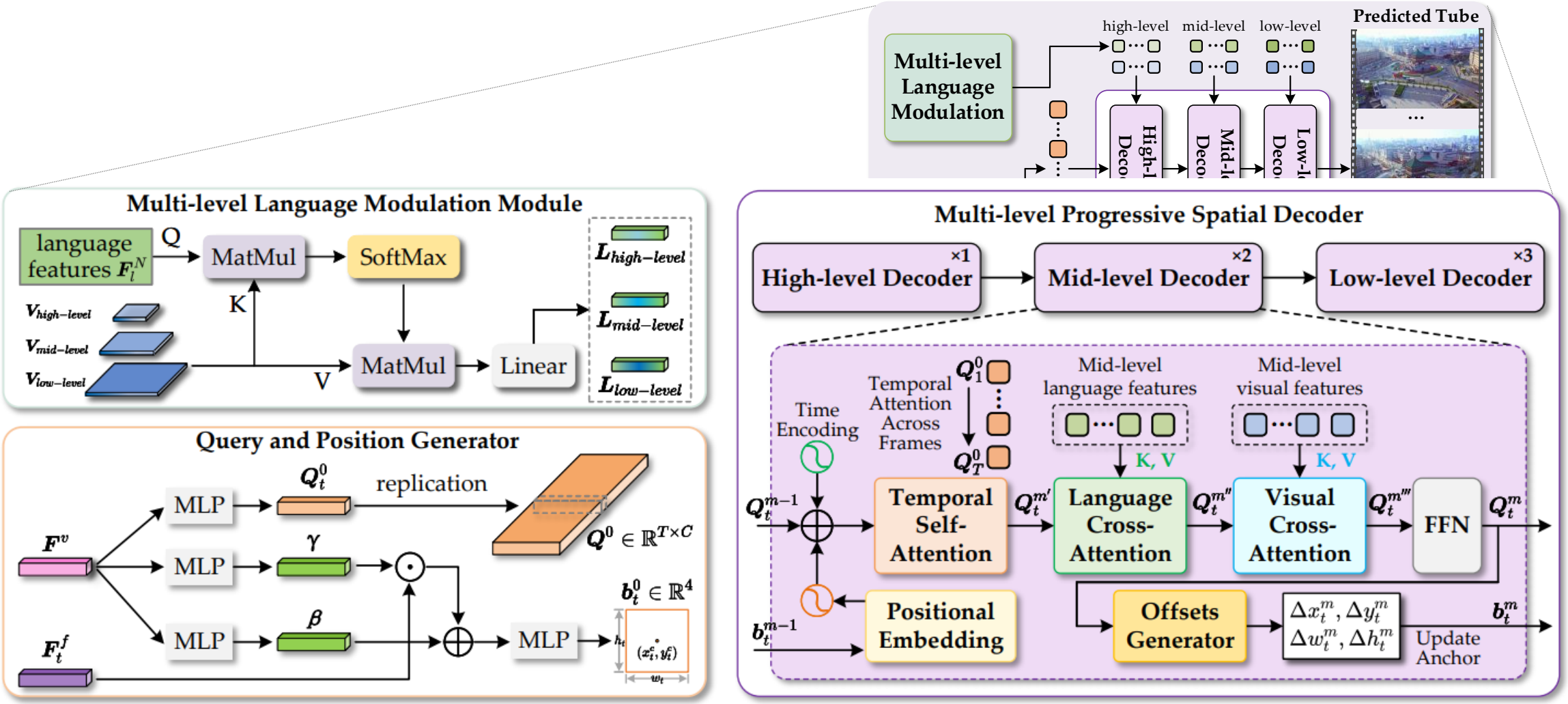
Method

Overview of SAVG-DETR: Finally, the Hierarchical Progressive Decoder utilizes multi-level language-vision features to guide queries to decode more relevant spatial information.



Method

Overview of SAVG-DETR: Finally, the Hierarchical Progressive Decoder utilizes multi-level language-vision features to guide queries to decode more relevant spatial information.



Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

· **Sufficient benchmarks based on state-of-the-art video grounding methods**

Methods	Visual Encoder	Language Encoder	m_vIoU	vIoU@0.3	vIoU@0.5	m_fAcc	fAcc@0.3	fAcc@0.5
Co-grounding [35]$_{\text{CVPR'21}}$	Darknet53	Bi-LSTM	10.24	21.66	6.11	11.17	16.29	8.40
DCNet [36]$_{\text{ACMMM'22}}$	Darknet53	BERT	11.65	23.58	8.79	13.10	17.64	9.21
TubeDETR [17]$_{\text{CVPR'22}}$	ResNet101	RoBERTa	22.60	32.91	20.49	23.84	29.69	22.00
STCAT[18]$_{\text{NeurIPS'22}}$	ResNet101	RoBERTa	<u>24.14</u>	<u>35.51</u>	<u>22.48</u>	<u>27.17</u>	<u>33.39</u>	<u>25.36</u>
SGFDN [45]$_{\text{ACMMM'23}}$	ResNet101	RoBERTa	20.13	28.16	15.47	19.13	22.71	17.39
CG-STVG [19]$_{\text{CVPR'24}}$	ResNet101	RoBERTa	21.23	28.82	19.04	22.32	26.24	20.41
VideoGrounding-DINO [20]$_{\text{CVPR'24}}$	Swin-Trans.	BERT	23.83	33.84	19.92	25.80	31.72	23.00
SAVG-DETR (Ours)	ResNet101	RoBERTa	**27.15**	**38.18**	**22.85**	**28.82**	**35.85**	**26.55**

- Our SAVG-DETR outperforms the state-of-the-arts consistently in all evaluation metrics.
- Co-grounding and DCNet are unable to handle T-frame global video features and lack spatio-temporal context modeling.
- We consistently outperform CG-STVG significantly despite that CG-STVG mines relevant and beneficial instance context during decoding, which shows the great potential of our method.
- In summary, the challenging SAVG task requires more delicate architectures or targeted improvements, and relatively generic improvements cannot significantly boost performance.

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

· **Ablation of key components of SAVG-DETR**

Encoder		Decoder			m_vIoU	m_fAcc
IMIB	CVFB	MLMM	QPG	MPSD	(%)	(%)
✓					22.38	25.46
✓				✓	19.37	21.84
✓	✓			✓	25.44	26.54
✓	✓	✓		✓	26.88	27.92
✓	✓	✓	✓	✓	**27.15**	**28.82**

intra-scale multi-modality interaction branch (IMIB), cross-scale visual-only fusion branch (CVFB), Multi-level Language Modulation Module (MLMM), query and position generator (QPG), multi-level progressive spatial decoder (MPSD).

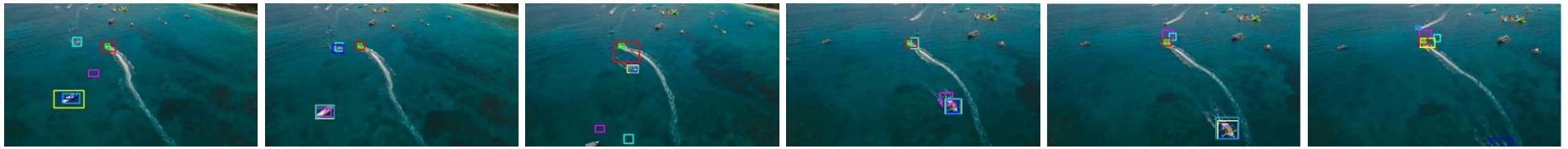
- The first row performs SAVG with only IMIB, vanilla decoder, and prediction head.
- In the second row, we further introduce multi-scale visual features into the decoder.
- In the third row, we further add the CVFB to achieve multi-scale fusion.
- The fourth row modulates language features through multi-scale visual features and guides object queries to capture spatial information more accurately in the decoder.
- The last row shows that after applying the QPG, our full-fledged model achieves the best performance.

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

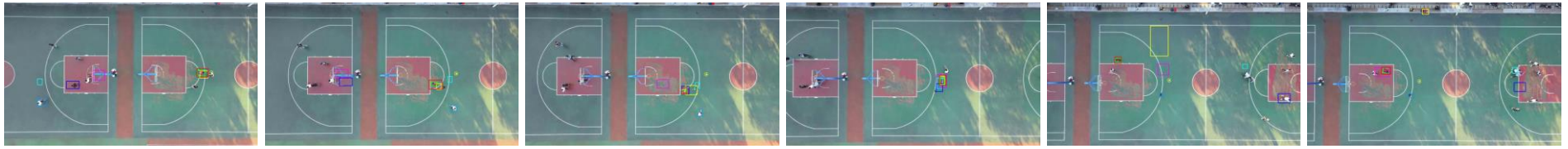
Qualitative results of different methods on the UAV-SVG benchmark. Our proposed SAVG-DETR (red) performs well and achieves reasonable localization results.
Our SAVG-DETR (red) can provide stable tracking based on early results with high prediction consistency across frames.

Expression 1: The only white speedboat towards the upper left of the sea, sailing with a rubber dinghy.

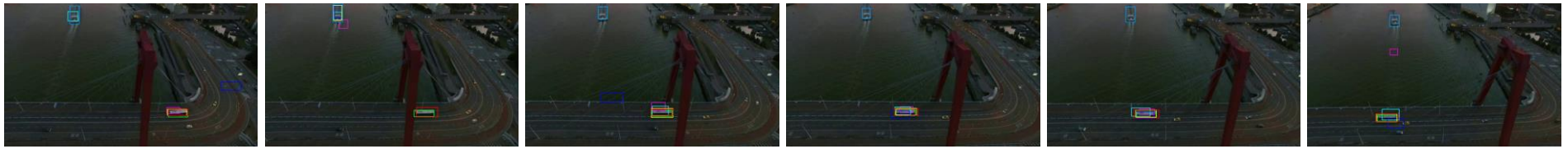


Expression 2: The only white coach that is driving near the building in the lower right corner, towards the left.

Expression 3: The boy wearing black pants and grey sneakers who is playing basketball on the right half of the court and first one counting from the top.



Expression 4: The white single-decker bus driving on the bridge, which is smaller than the ships in the water.



Expression 5: The red container truck driving on the leftmost lane of the highway on the far right side of the video.

Ground Truth SAVG-DETR TubeDETR STCAT SGFDN CG-STVG VideoGrounding-DINO

Where Does It Exist from the Low-Altitude: Spatial Aerial Video Grounding

Thank you for listening!
Welcome to our poster!

(Contact: zhanyangnwpu@gmail.com)