2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MIS: Light Response Agent for Video Comment with Multimodal Informative Seeking

Dong Zhang*[1,2], Tongfei Shen*[1], Zhiyu Tang[3], Shoushan Li†[1], and Guodong Zhou[1]

***Abstract*— Automatic response generation of video comments (RGVC) aims to generate a target reply to the content of the target comment based on the video context. Existing works for RGVC normally rely on large language models (LLMs), and mostly neglect the importance of extracting key information from both linguistic and visual perspectives. This limitation hinders the deployment of fluent and targeted response generation systems in real-world robotic and automated applications, where computational efficiency and precision are essential. In this work, we introduce a lightweight response agent with a novel multimodal informative seeking approach (MIS), which includes a Comment Context Retrieval (CCR) module and a Key Vision Selection (KVS) module to simultaneously seek essential information from both textual and visual modalities. Specifically, the CCR module enriches the dialogue context by retrieving relevant comments from other comment blocks, while the KVS module utilizes a spatial-temporal Transformer with cross-modal attention to highlight the most crucial information in the video. Moreover, we also build a large-scale user-level multimodal chitchat (UMC) dataset with exact comment-response interactions to better investigate RGVC. Extensive experiments demonstrate that our model effectively captures human points of interest and generates more fluent and diverse responses than state-of-the-art methods in both open and closed resources. These attributes make MIS particularly suitable for deployment in social robots, service automation, and other interactive robotic systems requiring real-time visual and linguistic inference.**

I. INTRODUCTION

Automatic response generation of video comments (RGVC) aims to generate replies to target comments based on video context. This task holds significant application value in human-robot interaction [1], [2], social companion systems [3], [4], interactive marketing [5], [6], and service robotics [7], [8]. Recently, [9] proposed a benchmark dataset and a primary response agent but still faces several challenges:

First, from **textual perspective**, the previous study normally uses an external general knowledge base to enhance text representation capabilities [10]. However, such a large-scale knowledge base with only general facts are often less relevant to the RGVC task. Therefore, task-relevant knowledge should be employed to augment textual representations. **Second**, from **visual perspective**, the previous study usually uses the entire video content as context for response generation. Comments typically address specific parts of a video, involving multiple topics [11], [12]. For example, in Figure 1, one comment *C1* focuses on the basketball game rules in the video, while another comment *C2* inquires about sneakers. Thus, using the entire video content can lead to unnecessary computational costs and noise towards the response generation of one specific comment. Therefore, key chunks of videos and specific frame regions should be selected. **Third**, from **dataset perspective**, the previous benchmark dataset has vague correspondences between comments and replies. For example, in Figure 1, *R2* responds to R1 rather than the initial comment *C1*, yet pairs like (*C1*, *R2*), (*C1*, *R3*) and even (*C1*+*C2*, *R1*) are included in the prior dataset, which complicates the model design for response generation. Therefore, exact matched comment-response pairs should be collected through meticulous human verification. **Finally**, from **real application perspective**, existing approaches normally rely on large language models (LLMs) [13], [14], which are computationally intensive and challenging to deploy in embedded systems or mobile robotic platforms that require low latency and high efficiency.

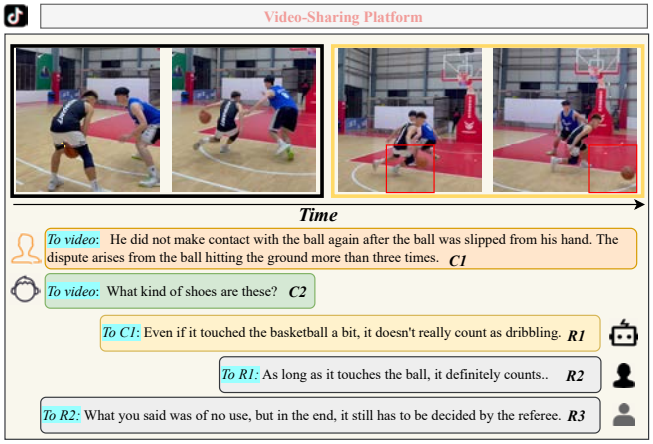

Fig. 1. *An example of video comments and responses in real-world scenario.* In the video, two basketball players engage in a one-on-one basketball game with a strict three-dribble limit per possession. They are disputing the number of dribbles during a specific round. Below the video, different users post their opinions or reply to specific users' comments in different video blocks.

To address these challenges, we propose a lightweight response agent for video comments with both textual retrieval and visual selection, termed multimodal informative seeking (MIS). This is designed to be efficient and scalable, making it suitable for robotic and automation applications. Additionally, we collect a more accurate comment-response matching dataset for RGVC evaluation, user-level multimodal chitchat

*Dong Zhang and Tongfei Shen are equal contributions. †Shoushan Li is corresponding author.

This work was supported by Jiangsu Province Frontier Program Project (BF2025036), Hong Kong RGC grant GRF 15611021, and NSFC grant (No. 62376178).

[1]School of Computer Science Technology, NLP Lab, Soochow University, China. dzhang@suda.edu.cn
[2]Jiangsu Key Lab of Language Computing, Suzhou.
[3]University of Queenland.

(UMC). Specifically, our proposed agent involves: *Feature Processing*: both textual and visual feature sequences are extracted by light CLIP [15]. *Textual Retrieval*: a retrieval strategy for the textual context (historical comment-response pairs of a video) is proposed to capture pairs closely related to the target comment, enhancing the representation of the target comment. *Visual Selection*: Using the target comment as a query, we leverage cross-modal attention to select key visual information (video clips and frame regions). *Fusion and Response Generation*: We fuse textual and visual representations through a lightweight decoder to generate the target response. In summary, our contributions are as follows:

- We propose MIS, the first lightweight agent to simultaneously seek informative parts of both textual and visual modalities for RGVC, making it suitable for robotic and automated interaction systems.
- We construct and manually verify a new RGVC dataset, user-level multimodal chitchat (UMC), ensuring fully corresponding comments and responses and supporting further development of human-robot interaction systems.
- We conduct systematic experiments and extensive analyses on both previous and our datasets, validating the effectiveness of response agent MIS. The key code fragments and data examples are available in the anonymous project: https://anonymous.4open.science/r/Mis-40F8/README.md.

II. Related Work

Automatic Response generation for video comments (RGVC) holds significant importance for enhancing interactive capabilities in robotic and automated systems [16], [17], [18]. To our knowledge, [9] is the first to focus on RGVC, but its limitations are noted in the introduction. Thus, we cover relevant studies involving video and sentence inputs expecting a response output.

A. Video Question Answering (VQA).

VQA is categorized into multiple-choice [19], [20] and open-ended question answering [21]. The former simplifies VQA into classifying answer options, while the latter generates answers. The representative model [21] uses optical flow to select frames as visual context, combined with question sentences for BLIP-based answer generation, which normally cannot be applied into most data and time-consuming. Unlike these, we adopt a hierarchical visual selection strategy to refine visual context and retrieve relevant comment-reply units from historical comments, enriching text semantics. This not only reduces computational overhead but also enhances semantic alignment—an essential feature for real-time robotic applications.

B. Multimodal Dialogue Generation (MDG).

To simulate real life and achieve MDG, various models are proposed to tackle multimodal dialogue challenges: Maria [22] uses UniLM and object tags to enhance visual perception, BLIP-2 [23] aligns visual and textual modalities with Q-Former, and Video-LLaVA [24] combines video frames and linguistic features efficiently. Unlike these, we target video and one-on-one comment-response pair, not continuous mixed conversations with multiple people. Additionally, we avoid relying on resource-intensive large multimodal language models, aiming for more efficient and specific performance in RGVC. This aligns well with the requirements of embedded and robotic systems.

III. Methodology

In this section, we first define the RGVC task, and then describe our proposed MIS. Figure 2 provides an overview of our approach, which comprises four main components: input feature processing, comment context retrieval (CCR), key vision selection (KVS), cross-modal decoder.

Task Definition. The goal of RGVC is to generate the response R for a given video V and a textual comment C, formulated in an auto-regressive manner as follows:

$$P(R|V,C,\theta)=\prod_{i=1}^{l}P(w_i|w_{<i},V,C,\theta) \tag{1}$$

where l denotes the max length of the generated response R, w_i means the i-th word in R, and θ is the set of trainable parameters of the model.

A. Input Feature Processing

Visual Encoder. Given a video sample V, we first sample T frames from this video using the average sampling method, so that it can be represented as $V \in \mathbb{R}^{T\times H\times W\times N}$ with T frames, N channels, height H and width W. We divide the video into M chunks, each video chunk contains $L=\lceil T/M\rceil$ frames, where $\lceil\cdot\rceil$ denotes the ceiling function. For each frame I, we extract both the global features and region features with pre-trained vision-language Transformer, CLIP [15]. Specifically, we first extract the features of the entire frame as global features, the process can be formulated as: $X^{\text{Global}}=f_{\text{v}}(I)$, where I is the video frame, f_{v} is the visual extractor, and X^{Global} denotes the global features of frame I.

Subsequently, we split frame I into B patches, represented as $I=\{I^1,I^2,\ldots,I^B\}$, and extract features from each patch as region features, formulated as follows:

$$X^{\text{Region}}=[f_{\text{v}}(I^1);f_{\text{v}}(I^2);\ldots;f_{\text{v}}(I^B)] \tag{2}$$

where I^i is the i-th patch of frame I, $[\cdot;\cdot]$ denotes the concatenation function, and X^{Region} represents the region features of frame I.

Finally, we get the final features of frame I as follows:

$$X=[X^{\text{Global}};X^{\text{Region}}] \tag{3}$$

Therefore, the video features we obtain can be represented as $X_v\in\mathbb{R}^{M\times L\times(B+1)\times D}$, where D is the dimension of video features.

Textual Encoder. We use CLIP to encode the information from historical comment context. For each comment sentence $c_i=\{e_1,e_2,\ldots,e_{m_i}\}$, we concatenate them with $[SEP]$ and format them as $C=[c_1,[SEP],c_2,[SEP],\ldots,c_n]$, where n denotes the number of historical context, m_i is the token length of sentence c_i. We then feed the concatenated comment

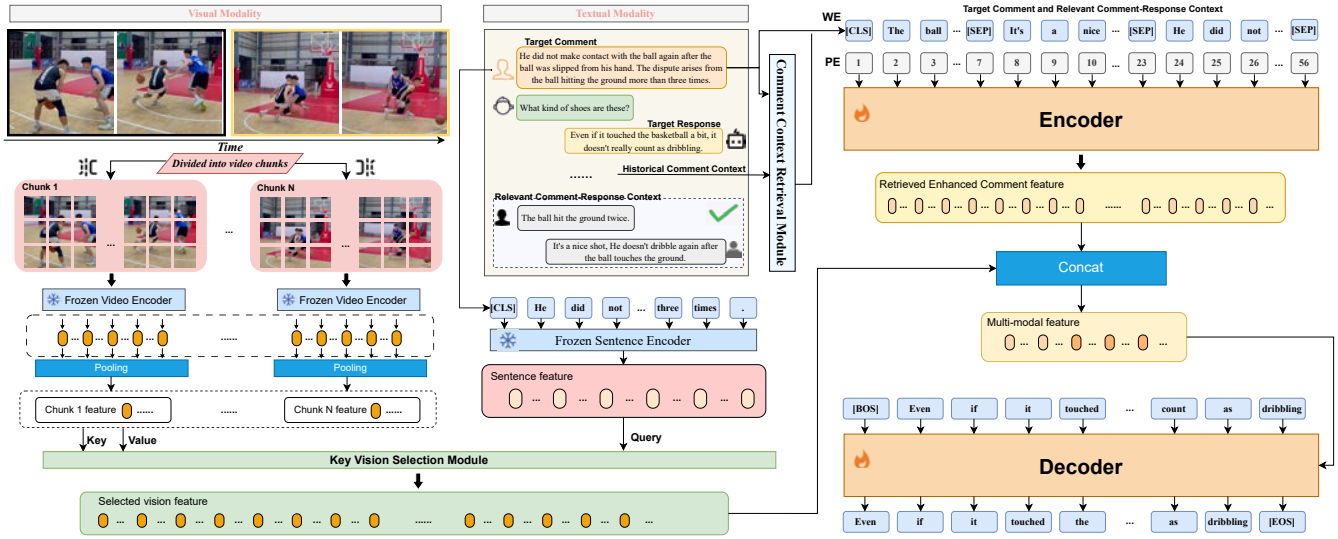

Fig. 2. *The overall architecture of our proposed* MIS. The model's workflow is structured into four steps: First, we segment the video into multiple chunks and utilize a frozen vision-language Transformer to extract features for each frame. Second, we actively enhance the semantics of the historical context with a comment context retrieval module. Third, a key vision selection module is employed to focus on vision areas that are most relevant to the target comment. Finally, we deploy an auto-regressive decoder to formulate the target response based on the fused features of vision and text.

context C into CLIP to obtain the text representation, which can be formulated as:

$$S = f_{\mathrm{t}}(C) \tag{4}$$

where f_{t} is the frozen textual encoder, S denotes the extracted text representation.

B. Comment Context Retrieval (CCR)

We believe that relying solely on the target comment information within a single comment block beneath a video is insufficient for comprehensive video understanding. Therefore, the objective of the comment context retrieval module is to filter out irrelevant historical comment-response blocks and construct a smaller candidate set for target comment, thereby enhancing the conversational context.

During the **training process**, we assume that different comment blocks under the same video share the same context. Consequently, for each video comment, we retrieve various video comment blocks under the same video in the training data by the nearest-neighbor algorithm, forming an external comment-reply pair H. We define $O = \{o_1, o_2, \ldots, o_n\}$ as the set of comment contexts from other comment blocks under the same video, where n is the number of different blocks. To ensure the relevance of the retrieved dialogues to the current target comment, we introduce a threshold, τ, to determine whether the current sample requires contextual semantic enhancement. Comment Context Retrieval on the training process can be formulated as follows:

$$v_i = \textsc{Strans}(o_i), Q = \textsc{Strans}(q_{train}) \tag{5}$$

$$o^* = \begin{cases} \arg\max_{v_i \in V_{train}} (v_i \cdot Q) & \text{if } \max_{v_i \in V_{train}} (v_i \cdot Q) \geq \tau \\ \text{None} & otherwise \end{cases} \tag{6}$$

where $\textsc{Strans}$ denotes the Sentence Transformer, q_{train} is the current target comment, and $V_{train} = \{v_1, v_2, \ldots, v_m\}$ is the embedding set of all comment contexts from other blocks under the same video.

Subsequently, we match the index o^* with the highest similarity to get the corresponding comment-reply pair H and concatenate H with C to obtain the semantically enriched target comment context $\hat{C} = [H; C]$, where C is the initial comment.

Since the different comment blocks under the same video are unseen during the **testing process**, we also utilize the nearest-neighbor algorithm to search the embedding space for the Top-1 comment-reply pairs in the training data that are close to the target comment of the test sample. We define $Z = \{z_1, z_2, \ldots, z_n\}$ as the set of different comment contexts from training data, where n is the number of different comment blocks. Comment Context Retrieval on the testing process can be formulated as follows:

$$v_i = \textsc{Strans}(z_i), Q = \textsc{Strans}(q_{test}) \tag{7}$$

$$z^* = \begin{cases} \arg\max_{v_i \in V_{test}} (v_i \cdot Q) & \text{if } \max_{v_i \in V_{test}} (v_i \cdot Q) \geq \tau \\ \text{None} & otherwise \end{cases} \tag{8}$$

where $V_{test} = \{v_1, v_2, \ldots, v_n\}$ is the embedding set of all comment contexts from training data.

After obtaining the semantically enriched comment $\hat{C}$, we encode it using a trainable textual encoder: $E = \textsc{Enc}_{\mathrm{txt}}(\hat{C})$, where $\textsc{Enc}_{\mathrm{txt}}$ is the trainable textual encoder, and E denotes the representation of semantically enriched comment, with a dimensionality of D_2.

C. Key Vision Selection (KVS)

Chunk Selection (CS). Given the video features $X_v \in \mathbb{R}^{M \times L \times (B+1) \times D}$, the goal of chunk selection is to identify

the video chunks most relevant to target comments. Firstly, we separate the global and regional features of the video, and then perform spatial pooling on the global features to derive the chunk features of the video. Specifically, let f_l^m indicate the l-th frame feature in m-th chunk, and then obtain the chunk features by performing spatial pooling on frame features. The process can be formulated as follows:

$$f^m = \text{Pooling}(f_1^m, f_2^m, \ldots, f_L^m) \tag{9}$$

where f^m is the m-th chunk features, and $\text{Pooling}(\cdot, \cdot)$ denotes the mean pooling function here.

Given the chunk features $F = [f^1; f^2; \ldots; f^M]$, region features X^{Region}, historical context features S, and semantically enriched historical context features E, we initially apply cross-attention between the chunk features F and initial target comment features S. Subsequently, we implement a differentiable top-k feature selection process on X^{Region}, which can be formulated as follows:

$$Q = \phi_s(S), K = \phi_f(F), V = X^{\text{Region}} \tag{10}$$

$$X_{\text{cs}} = \underset{\text{Top}_k}{\text{SELECTOR}}(\text{softmax}(\frac{QK^T}{\sqrt{d_k}}), V) \tag{11}$$

where ϕ_s and ϕ_f are fully connected layers, Selector is a differentiable top-k selection function implemented with the Gumbel-Softmax trick [25]. X_{cs} represents the chunk-selected features resampled from multiple video blocks, which can be described as $X_{\text{cs}} \in \mathbb{R}^{Top_k \times L \times B \times D}$.

Region Selection (RS). For the l-th frame within the chunk-selected features X_{cs}, which comprises B patch-level region features, we aim to further select the region features most relevant to the initial comment features S, the process is formulated as follows:

$$Q = \phi_{s'}(S), \quad K = \phi_x(X_{\text{cs}}), \quad V = X_{\text{cs}} \tag{12}$$

$$X_{\text{rs}} = \underset{\text{Top}_j}{\text{SELECTOR}}(\text{softmax}(\frac{QK^T}{\sqrt{d_k}}), V) \tag{13}$$

where $\phi_{s'}$ and ϕ_x are fully-connected layers. X_{rs} represents the region-selected features resampled from multiple frame patches and can be expressed as $X_{\text{rs}} \in \mathbb{R}^{Top_k \times L \times Top_j \times D}$.

We then flatten the region-selected features and concatenate them with the chunk features F, and obtain the final visual representation X_{vis} by projecting concatenated visual features into text token space.

$$X_{\text{cat}} = [F; X_{\text{rs}}], X_{\text{vis}} = \phi_f(X_{\text{cat}}) \tag{14}$$

D. Cross Modal Decoder

Our model is based on the Encoder-Decoder paradigm, where it fuses semantically enhanced textual features with selected key vision features, formulated as:

$$X_{\text{src}} = [E; X_{\text{vis}}] \tag{15}$$

In this setup, the target response is defined as $Y = \{y_1, y_2, \ldots, y_n\}$. The cross-modal decoder then generates the responses in an auto-regressive manner. We optimize the model using the cross-entropy loss as:

$$L_{\text{gen}} = -\mathbb{E}_{y_i \sim Y} logp(y_i | y_{<i}, X_{\text{src}}) \tag{16}$$

IV. Experimentation

A. Experimental Setting

Dataset. We evaluate models on the TikTalk dataset [9], which comprises 38K videos and 367K conversations sourced from DouYin. The dataset is divided into training, validation, and test sets, containing 35 703, 1000, and 2000 videos respectively. Following [9], we eliminate the videos with corrupted formats and finally obtain the dialogue counts: 418 341 for the training set, 1402 for the validation set, and 2827 for the test set.

For our UMC dataset, we collect 22 766 videos from DouYin[1] and extract a total of 307 000 user-level conversations from various video comment blocks. As the same with [9], we ensure each conversation in the validation and test sets includes at least five ground-truth responses. The dataset is divided into 300 000 training samples, 2000 validation samples, and 5000 test samples, respectively.

Metrics. We evaluate our models using both automatic metrics and human evaluation.

For automatic metrics. We utilize BLEU [26], ROUGE [27] and CIDEr [28] to assess the similarity and relevance of generated responses to reference responses. Additionally, we adopt Distinct [29] to measure the diversity of generated responses. We compute the metrics individually for each ground-truth response and then average the results to obtain the final metrics for each conversation. We adopt the public NLG evaluation code[2] to calculate these metrics.

For human evaluation. Following [9], we define three distinct dimensions to evaluate generated responses: 1) Sensibleness (*assessing whether the model-generated responses are fluent and diverse, avoiding simple phrases like "me too" and ensuring they are not mere imitations of previous comments*), 2) empathy (*evaluating whether the model-generated responses resonate empathetically with humans and are not easily discerned as AI-generated responses.*) and 3) the overall quality based on the first two aspects.

We randomly sample 100 instances from the test set of the TikTalk dataset to form a questionnaire and invite 25 annotators to conduct manual evaluations according to the above three criteria on a crowdsourcing platform. After that, we weight the annotators' rankings of the quality of model responses and scale the scores of each aspect to 5.

B. Baselines and Implementation Details

We conduct experiments with ten competitive models, including three textual modal models and seven multimodal models.

For Textual Baselines. Following [9], we adopt DialoGPT [30] and ChatGLM-6B [31] as textual baselines and additionally introduce Qwen1.5-Chat models with different parameter size [32].

For Multimodal Baselines. In addition to Livebot [33], BLIP-2 [23], and the *SOTA* Maria+C^{3}KG [9] established by the TikTalk benchmark, we also introduce Mist (*A*

[1]https://www.douyin.com

[2]https://github.com/Maluuba/nlg-eval

Type	Model	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	Dist-1	Dist-2	Dist-3
Textual	DialoGPT [30]	1.78	1.01	0.35	8.62	12.76	**5.03**	24.89	45.35
	ChatGLM-6B [31]	1.32	0.48	0.15	7.24	9.04	4.17	27.77	50.22
	Qwen1.5-1.8B-Chat [32]	3.69	1.88	0.90	11.63	21.25	3.75	26.85	45.64
	Qwen1.5-7B-Chat [32]	3.94	2.03	0.96	12.16	23.16	4.26	29.94	50.85
Multimodal	Livebot [33]	2.15	1.16	0.35	9.88	13.68	4.44	21.41	39.34
	BLIP-2_Img* [23]	3.42	-	0.87	8.34	12.05	-	**38.40**	64.63
	BLIP-2_Video* [23]	3.41	-	0.75	11.60	26.19	-	37.32	62.44
	MIST [20]	2.48	0.90	0.35	10.66	9.89	2.48	18.63	36.28
	Qwen2-VL^ [34]	1.95	0.54	0.18	8.41	4.17	1.65	22.42	47.74
	GPT-4o^ [35]	2.78	0.92	0.36	10.08	6.90	2.95	35.20	**65.63**
	Maria+C³KG [9]	4.68	1.95	0.88	13.69	20.47	2.81	22.78	43.54
	Maria+CCR	4.79	2.24	1.13	13.26	19.79	2.76	24.41	47.31
	MIS (Ours)	**5.37**	**2.68**	**1.34**	**14.11**	**29.85**	4.26	36.13	62.02
Ablations	MIS (w/o Video)	5.01	2.54	1.27	13.61	29.12	4.53	35.91	60.86
	MIS (w/o CCR)	5.17	2.57	1.24	14.01	27.79	4.13	34.86	60.75
	MIS (w/o CS)	5.18	2.54	1.23	13.85	28.88	4.18	34.08	57.88
	MIS (w/o RS)	4.96	2.44	1.14	13.54	26.78	3.87	31.29	53.92
	MIS (w/o CS & RS)	3.88	1.73	0.89	11.66	21.86	2.87	28.06	51.53

TABLE I

THE PERFORMANCE COMPARISON OF DIFFERENT BASELINES AND OUR APPROACH ON TIKTALK DATASET. RESULTS MARKED WITH AN ASTERISK (*) ARE TAKEN FROM [9]. MODELS MARKED WITH A CARET (^) REPRESENT ZERO-SHOT SETTING.

multi-modal iterative spatial-temporal transformer framework for long-form video question answering [20]), Qwen2-VL (*A large multimodal model capable of dynamic resolution processing and long video understanding* [34]), and GPT-4o (*about 200B parameters, as the* SOTA *in closed source models* [35]) as multimodal baseline models. Furthermore, we design a variant of Maria+C^{3}KG, called Maria+CCR, to test the effectiveness of our CCR.

For our approach, the historical context representation and visual representation for the vision selection module are extracted by the ViT-B/16 version of pre-trained ChineseCLip. Other hyperparameters can refer to Appendix.

C. Main Results

We conduct comprehensive experiments in the TikTalk dataset. Due to the limit of computing resources and time, we select the top four well-performing models and their variants to conduct experiments on our proposed dataset UMC.

On TikTalk Dataset. As depicted in Table I, we can find that each baseline has an advantage in a small number of metrics, but does not exhibit a clear pattern of advantages. Among them, in most cases, due to extremely large parameter size, text-based large language models (LLMs, like **Qwen**) show superior performance than prior multimodal models (e.g., **Livebot, MIST**). To our surprise, the mainstream multimodal large language model (**Qwen2-VL**) struggles to grasp the key points in videos and tend to merely echo the user's comments. Although **GPT-4o** has a larger scale, it performs poorly due to the inability to fine-tune the closed source model. Compared to generation models in various multimodal scenarios that have emerged recently, **Maria+C^{3}KG** as SOTA for RGVC performs poorly on many cases, however, it still outperforms other baselines on common metrics: BLEU-2 and ROUGE-L. The above phenomenon indicates that it is necessary to design a new approach for RGVC to improve response generation quality in various metrics. Through the results in Table I, we observe that our model **MIS** surpasses all mainstream baselines (**Livebot, Maria+C^{3}KG**) and even LLMs (**Qwen, BLIP-2-7B, Qwen2-VL-72B GPT-4o-200B**) in all metrics except Dist. While our proposed agent built on BART only has a parameter count of only 0.1B. This is mainly due to our proposed informative seeking strategies on different modalities simultaneously. Therefore, *the proposed agent Mis possesses a lightweight nature, enabling it to achieve the intended performance objectives with extremely low resource overhead. Accordingly, the agent exhibits remarkable ease of deployment across multiple scenarios, and is particularly suitable for resource-constrained mobile scenarios.*

Additionally, in terms of diversity metrics (*Dist*), this is only a small aspect of evaluating the quality of responses. As is well known, LLMs can generate diverse responses due to their large number of parameters and large scale of pre-trained corpus. Especially like **BLIP-2**, which incorporates a large amount of video information during pre-training. But they did not perform the best on other metrics. So just because the *Dist* metric is outstanding, the generation quality may not be good. For specific examples, please refer to the case study.

On our UMC Dataset. To assess the robustness and reliability of our **MIS**, we conduct additional experiments using our UMC dataset. This new dataset includes *5000* user-level test samples, as opposed to only *2827* non-user-level test samples in the TikTalk dataset, which makes our experimental results more convincing. As shown in Table II, our proposed **MIS** outperforms other strong baselines on all metrics. We additionally replace the knowledge graph component in Maria+C^{3}KG with our proposed retrieval-based CCR module. As a result, we observe improvements in both similarity and diversity metrics for this variant **Maria+CCR**

Type	Model	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	Dist-1	Dist-2	Dist-3
Textual	Qwen1.5-7B-Chat [32]	4.32	2.47	1.58	11.64	32.71	2.80	20.12	33.71
Multimodal	BLIP-2_Img [23]	1.95	0.82	0.43	6.77	5.96	1.72	26.16	50.55
	BLIP-2_Video [23]	2.13	0.90	0.46	7.26	7.33	1.81	26.97	52.46
	Maria+C³KG [9]	4.61	2.18	1.34	12.74	24.64	1.96	18.57	37.60
	Maria+CCR	4.70	2.49	1.56	12.76	30.51	2.79	24.00	44.13
	MIS (Ours)	**5.34**	**2.87**	**1.86**	**13.36**	**34.53**	**3.17**	**31.81**	**58.28**
Ablations	MIS (w/o Video)	4.91	2.56	1.61	12.78	31.38	3.10	31.24	57.40
	MIS (w/o CCR)	5.32	2.81	1.75	13.30	32.84	2.99	31.07	57.60
	MIS (w/o CS)	5.10	2.69	1.69	12.95	32.23	2.95	29.11	53.65
	MIS (w/o RS)	5.00	2.66	1.68	12.85	31.89	3.03	30.96	57.95
	MIS (w/o CS & RS)	4.92	2.59	1.61	12.69	30.37	2.82	28.72	53.40

TABLE II

THE PERFORMANCE COMPARISON OF THE TOP FOUR MODELS AND THEIR VARIANTS ON OUR UMC DATASET.

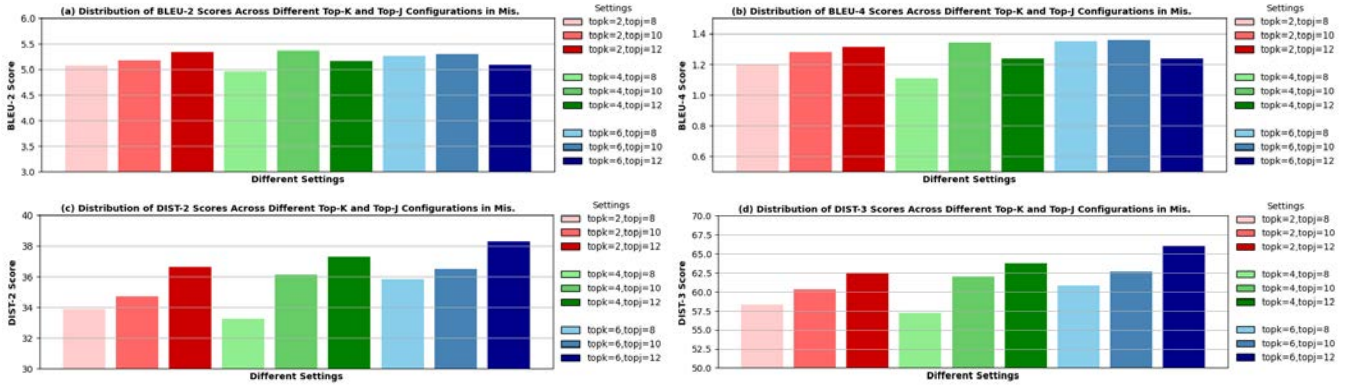

Fig. 3. Performance of Similarity and Diversity Metrics Across Different Top-K and Top-J Settings.

Model	Perspective		
	Sensibleness	Empathy	Overall
Qwen1.5-7B-Chat	0.61	0.62	0.68
Maria+C³KG	0.51	0.52	0.51
GPT-4o	**1.29**	1.16	**1.20**
MIS (Ours)	1.16	**1.19**	1.15

TABLE III

PERFORMANCE OF THE TOP FOUR MODELS IN HUMAN EVALUATION.

compared to Maria+C^{3}KG. Additionally, experimental results on diversity metrics demonstrate that our method not only generates more diverse responses but also proves to be more robust compared to other models when using user-level comment-response matching dataset.

D. Analysis and Discussion

Ablation Study. We evaluate the performance of **MIS** by removing video modality (**w/o Video**), removing CCR only (**w/o CCR**), removing chunk selection only (**w/o CS**), removing region selection only (**w/o RS**), and removing both CS and RS (**w/o CS&RS**, i.e., without KVS). The results of our ablated approaches are presented in the last five rows of Table I and II. From these tables, we can find that discarding any module in our approach results in varying degrees of performance degradation. Interestingly, removing KVS (**w/o CS&RS**) results in higher performance loss compared to removing all visual features (**w/o Video**) regarding the ROUGE-L and CIDEr metrics on both datasets. This precisely indicates that using visual features directly will bring apparent noise, indicating the necessity of designing our KVS. This also confirms our motivation in our introduction.

Human Evaluation. For a fair comparison, we mainly conduct human evaluation on the public dataset TikTalk. From the results depicted in Table III, we can observe that: 1) Relying on its strong memory and expressive capabilities (parameters) to handle diverse and noisy user comments, GPT-4o achieves the highest overall quality in response generation. This also suggests that due to the divergent nature of GPT-4o's responses, its automatic evaluation scores based on n-gram matching (BLEU) are relatively low, highlighting the necessity of reasonable manual evaluations. 2) From an empathy perspective, our proposed Mis obtain the highest scores, possibly because GPT-4o still tends to generate objective responses, making it easier to label as "AI-generated". 3) Regarding all aspects of human evaluation, our Mis performs much better than SOTA Maria+C^{3}KG and 7B LLM Qwen, almost comparable to GPT-4o, which reveals the excellent ability of our Mis for RGVC task.

Computational Complexity. We assume T frames are extracted, with BLIP-2 using 32 tokens for visual selection. BLIP-2 extracts image features as $N \times D_1$, where N is the

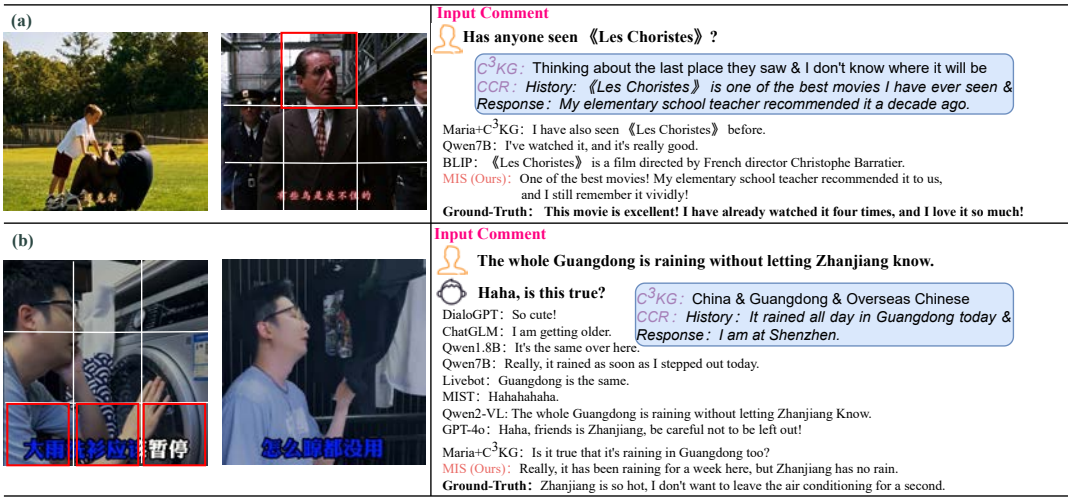

Fig. 4. Two cases in (a) UMC and (b) TikTalk datasets, predicted by different models, respectively. We only provide two chunks and several regions with red boxes that our MIS focuses on for each case on the left due to the space limit. *C^3KG* denotes the relevant knowledge used by SOTA Maria+C^{3}KG and *CCR* denotes the retrieved relevant context by our MIS.

patch number and D_1 is the corresponding dimension in BLIP-2. Consistent with the symbols used in this paper, MIS selects TopK among M video chunks and TopJ among K patches, where each video chunk contains L frames. The image features in MIS correspond to dimension D, where $D < D_1$. **For BLIP-2**, it uses 32 randomly initialized tokens to perform cross-attention queries on image features, known as Q-Former. Its time complexity is $O(T\times 64\times N\times D_1)$. **For MIS**, its visual selection stage is divided into chunk selection and region selection stages. In chunk selection, its selection time complexity is $O(2\times TopK\times T\times B\times D)$. In region selection, its selection time complexity is $O(2\times TopK\times TopJ\times L\times B\times D)$. Therefore, the total time complexity of MIS's selection process is $O((T+L\times TopJ)\times(2\times TopK\times B\times D))$, where $(N\times D_1)\approx 8\times(2\times TopK\times B\times D)$ and $(T\times 64)\approx 20\times(T+L\times TopJ)$, which is lower than BLIP-2. Furthermore, with a parameter count of only 0.1 billion, our agent is significantly smaller in scale compared to existing large language model (LLM)-based response agents (e.g., GPT-4o with 200B parameters). This indicates that our agent is lightweight, capable of achieving the desired performance with minimal resource consumption. Thus, it can be easily deployed across various scenarios, particularly on mobile devices.

Effect of Different K and J in KVS. As illustrated in Figure 3, we report the different settings of the selected chunks (K) and regions (J) in our KVS module. From this figure, we can observe that setting a specific set of K (=4) and J (=10) can achieve balanced results, while other settings, although they can achieve better results in several cases, are not the best.

Case Study. We present two examples in Figure 4, including the two main visual chunks, several attracted regions with red boxes, input comments, and response results generated by mainstream competitive methods. From case (a), we can see that our **MIS** can effectively capture the face region in the second chunk, which is a famous movie star, and can help us generate movie-related responses. Besides, we also retrieved comments and replies related to the input comment, further enhancing the multimodal representation for RGVC. Compared to other strong baselines, our response is clearly more reasonable. Additionally, from case (b), we can observe that although multiple patches are visually noticed (three red boxes), resulting in less visual gain, we can still rely on comment context retrieval to enhance input information and improve response quality. The generated results precisely support our inference: our agent's auto response delves deeper into the comments and provides more detailed content, like "raining for a week" and "Zhanjiang". Therefore, even if the Dist metric of a model is high, it cannot indicate that its response aligns with human preferences.

V. Conclusion

We propose MIS, a automatic response agent for video comment, and build a user-level multimodal chitchat dataset UMC with matching comment-response pairs. Our MIS can simultaneously capture comment-relevant points in videos (KVS) and augment textual information with related historical context (CCR). Extensive experiments demonstrate the effectiveness of our MIS and the necessity of our collected UMC dataset.

References

[1] L. Li, D. Zhang, S. Zhu, S. Li, and G. Zhou, "Response generation in multi-modal dialogues with split pre-generation and cross-modal contrasting," *Inf. Process. Manag.*, vol. 61, no. 1, p. 103581, 2024.

[2] F. Kong, P. Wang, S. Feng, D. Wang, and Y. Zhang, "TIGER: A unified generative model framework for multimodal dialogue response generation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 2024, pp. 16 135–16 141.

[3] Z. Tian, Z. Xie, F. Lin, and Y. Song, "A multi-view meta-learning approach for multi-modal response generation," in *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, 2023, pp. 1938–1947.

[4] J. Huang, A. Wang, L. Gao, L. Song, and J. Su, "Response enhanced semi-supervised dialogue query generation," in *AAAI 2024*, M. J. Wooldridge, J. G. Dy, and S. Natarajan, Eds. AAAI Press, 2024, pp. 18 307–18 315.

[5] N. Akoury, Q. Yang, and M. Iyyer, "A framework for exploring player perceptions of llm-generated dialogue in commercial video games," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, 2023, pp. 2295–2311.

[6] W. Lajewska, "Grounded and transparent response generation for conversational information-seeking systems," in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, 2024, pp. 1142–1144.

[7] H. Zhang, A. Saood, J. J. G. Cardenas, X. Hei, and A. Tapus, ""oh! it's fun chatting with you!" a humor-aware social robot chat framework," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 5520–5526.

[8] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "Spatialbot: Precise spatial understanding with vision language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9490–9498.

[9] H. Lin, L. Ruan, W. Xia, P. Liu, J. Wen, Y. Xu, D. Hu, R. Song, W. X. Zhao, Q. Jin, and Z. Lu, "Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world," in *Proceedings of the 31st ACM International Conference on Multimedia, ACM MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 2023, pp. 1303–1313.

[10] J. Lee, S. Park, S. Park, H. Kim, and H. Kim, "A framework for vision-language warm-up tasks in multimodal dialogue models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 2023, pp. 2789–2799.

[11] Y. Wang, Z. Zheng, X. Zhao, J. Li, Y. Wang, and D. Zhao, "VSTAR: A video-grounded dialogue dataset for situated semantic understanding with scene and topic transitions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 2023, pp. 5036–5048.

[12] H. Liu, P. Wang, Z. Zhu, Y. Wang, and Y. Wang, "CE-VDG: counterfactual entropy-based bias reduction for video-grounded dialogue generation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, 2024, pp. 2958–2968.

[13] L. Grassi, Z. Hong, C. T. Recchiuto, and A. Sgorbissa, "Grounding conversational robots on vision through dense captioning and large language models," in *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 5492–5498. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10611232

[14] P. Li, L. Cao, X.-M. Wu, X. Yu, and R. Yang, "Ugotme: An embodied system for affective human-robot interaction," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 5542–5548.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, vol. 139, 2021, pp. 8748–8763.

[16] L. Ma, J. Li, M. Li, W. Zhang, and T. Liu, "Policy-driven knowledge selection and response generation for document-grounded dialogue," *ACM Trans. Inf. Syst.*, vol. 42, no. 2, pp. 49:1–49:29, 2024.

[17] F. Askari, H. Abdollahi, K. S. Haring, and M. H. Mahoor, "A reinforcement learning-based social robot for personalized learning in children with autism," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 12 445–12 450.

[18] A. Nardelli, A. Sgorbissa, and C. T. Recchiuto, "Personality- and memory-based software framework for human-robot interaction," in *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. IEEE, 2024, pp. 17 388–17 394. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10611168

[19] S. Buch, C. Eyzaguirre, A. Gaidon, J. Wu, L. Fei-Fei, and J. C. Niebles, "Revisiting the "video" in video-language understanding," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022, pp. 2907–2917.

[20] D. Gao, L. Zhou, L. Ji, L. Zhu, Y. Yang, and M. Z. Shou, "MIST : Multi-modal iterative spatial-temporal transformer for long-form video question answering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023, pp. 14 773–14 783.

[21] Y. Wang, Y. Wang, P. Wu, J. Liang, D. Zhao, and Z. Zheng, "LSTP: language-guided spatial-temporal prompt learning for long-form video-text understanding," *CoRR*, vol. abs/2402.16050, 2024.

[22] Z. Liang, H. Hu, C. Xu, C. Tao, X. Geng, Y. Chen, F. Liang, and D. Jiang, "Maria: A visual experience powered conversational agent," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 2021, pp. 5596–5611.

[23] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, vol. 202, 2023, pp. 19 730–19 742.

[24] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *CoRR*, vol. abs/2311.10122, 2023.

[25] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[26] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 2002, pp. 311–318.

[27] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[28] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 4566–4575.

[29] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016, San Diego California, USA, June 12-17, 2016*, 2016, pp. 110–119.

[30] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 270–278.

[31] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, and J. Tang, "GLM-130B: an open bilingual pre-trained model," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

[32] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[33] S. Ma, L. Cui, D. Dai, F. Wei, and X. Sun, "Livebot: Generating live video comments based on visual and textual contexts," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 2019, pp. 6810–6817.

[34] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *CoRR*, vol. abs/2409.12191, 2024.

[35] OpenAI, "Hello gpt-4o," 2024.