Coarse-to-Fine Multimodal Information Selection for Video Speaking Style Recognition with Large Language Models

Beibei Zhang[1], Yanan Lu[2], Fen Lin[2], Tongwei Ren[1]*
[1]State Key Laboratory for Novel Software Technology, Nanjing University [2]Tencent
zhangbb@smail.nju.edu.cn {yananlu,felicialin}@tencent.com rentw@nju.edu.cn

Abstract

Video Speaking Style Recognition (VSSR) aims to classify conversation videos into different types, significantly facilitating human interaction understanding. Recent approaches explore the potential of large language models (LLM) in VSSR with a training-free process. However, directly integrating all multimodal data yields suboptimal results, since the great redundancy in visual data can overshadow other valuable multimodal information, such as valuable textual dialogues and critical visual clues. To address this, we propose CFMiS (**C**oarse-to-**F**ine **M**ultimodal **I**nformation **S**election), a novel framework for VSSR that dynamically obtain valuable multimodal data via coarse-to-fine selection, enhancing LLM reasoning for VSSR. Specifically, the core of CFMiS are two cascaded modules: 1) a text-dominant modality selection module firstly selects VSSR-required modalities originating from text-based prediction; and 2) if vision is included in the selected modalities, a visual refinement module iteratively collects VSSR-relevant critical visual clues. The former resolves which modality to utilize, while the latter determines which information to adopt from selected modalities, efficiently alleviating information redundancy. Extensive experiments on multiple datasets prove that CFMiS is highly effective for VSSR, outperforming all existing training-free approaches and most training-based methods.

1 Introduction

Video speaking style recognition (VSSR) focuses on identifying various types of conversations within videos, providing an essential viewpoint for comprehending human interactions. It serves a significant role in numerous applications, including video content analysis (Wu and Krahenbuhl, 2021), conversation head generation (Zhou et al., 2022), and virtual agent design (Aneja et al., 2021). Consequently, VSSR has garnered more and more interest in recent years.

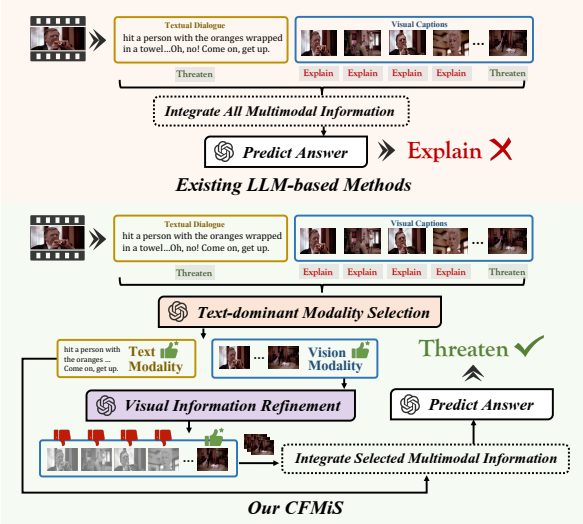

Figure 1: Comparison between existing LLM-based methods and our CFMiS for VSSR. Here, speaking style labels in gray background are predicted by separately inputting textual dialogue and each visual caption to LLM.

Existing training-based VSSR methods commit to designing various sophisticated networks to capture valuable clues for VSSR from vision-only (Wu and Krahenbuhl, 2021; Islam and Bertasius, 2022; Wang et al., 2023a; Li et al., 2024) or multimodal data (Sun et al., 2022; Zhang et al., 2023; He et al., 2024; Zhang et al., 2025). Nevertheless, VSSR is still a challenging task that falls short of the ideal (Strafforello et al., 2023), since it is a difficult fine-grained classification task with limited annotated training datasets. Confronting this issue, recently, several approaches (Bhattacharya et al., 2023; Lin et al., 2024) explore to leverage LLM, which exhibits remarkable capabilities in content understanding and reasoning (Qin et al., 2023; Zhang et al., 2024b; Sun et al., 2024b; Wang

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29322–29337

et al., 2025a), to comprehend the relationships between videos and speaking styles with a training-free process. Typically, these approaches leverage vision-language models (VLM) to caption densely sampled frames, thus converting the visual contents of videos into language. Visual captions are then integrated with textual dialogues to feed LLM, which subsequently reasons over these multimodal information to answer the provided VSSR query.

Despite the progress, immediately integrating all multimodal information leads to a suboptimal result for these methods. As shown in Figure 1, text modality always holds an advantage for VSSR because textual dialogue inherently contains rich semantic information highly related to the conversation. In contrast, visual contents, typically involving face-to-face talking shots, follow a similar pattern for most of the time in conversation videos, pointing to misleading neutral speaking styles, *e.g.,* Explain. Only few key behaviors that are indicative of the target speaking style occasionally intersperse throughout. The great redundancy in visual data ultimately overshadows valuable information relevant to the real speaking style during multimodal information integration (Zhang et al., 2023, 2025), mistaking LLM to generate an inaccurate answer.

Motivated by these observations, we propose a novel framework, CFMiS, dynamically conducting coarse-to-fine multimodal information selection to enhance LLM reasoning on VSSR. The target of CFMiS is to retain valuable multimodal information for VSSR, which are semantic-rich textual dialogues and sparse critical visual clues. Specifically, CFMiS designs two cascaded modules to chase down such information. Inspired by the self-evaluation module in LLM-based agents for gathering valid information (Gao et al., 2023; Wang et al., 2024), the first prime module of CFMiS is a text-dominant modality selection module, evaluating which modality is required for VSSR originating from the text-based prediction. We leverage LLM to implement the evaluation process with two steps, where the first step confirms the validity of textual information, in case of occasional detrimental contents within dialogues that leads to invalid prediction. The second step concludes the status of vision modality, evaluating if visual information can make additional contribution except for the existing text-based prediction. The dominant role of text in this module safeguards its prominent contribution to conversation understanding for VSSR.

The second module is a visual information refinement module, further determining which information of selected modalities is required for VSSR. CFMiS designs an iterative process to dynamically collect critical visual clues that are distinctly relevant to VSSR. To be specific, a video is firstly divided into multiple clips and subsequently fed to LLM to evaluate the relevance to VSSR. Visual captions of high-relevant clips are then iteratively gathered until the integration result is sufficient for LLM to generate an absolutely confident VSSR response. From the coarse-grained modality selection to fine-grained information selection, CFMiS efficiently gathers valuable multimodal information for VSSR, significantly alleviating the negative influence of information redundancy for multimodal fusion.

In general, the main contributions of this work can be summarized as follows:

- We propose a novel method, CFMiS, which conducts coarse-to-fine multimodal information selection to obtain valuable information for VSSR with large language models. The proposed CFMiS designs a text-dominant modality selection module, dynamically determining which modality is required for VSSR.
- To further capture valuable clues from selected modalities, CFMiS involves a visual information refinement module, iteratively collecting critical visual clues for VSSR.
- Extensive experiments and ablation studies are conducted to demonstrate that CFMiS is highly effective for VSSR, outperforming all training-free methods and most training-based methods.

2 Related Work

Video Speaking Style Recognition. VSSR aims to recognize human conversation types in videos (Wu and Krahenbuhl, 2021), whereas previous works concentrate on analyzing speaking styles in speeches (Wei et al., 2013, 2014). Existing VSSR methods fall into two categories: 1) Visual-focused approaches, which use sophisticated networks, such as transformer (Wu and Krahenbuhl, 2021; Xiao et al., 2022; Sameni et al., 2023), state space model (Islam and Bertasius, 2022;

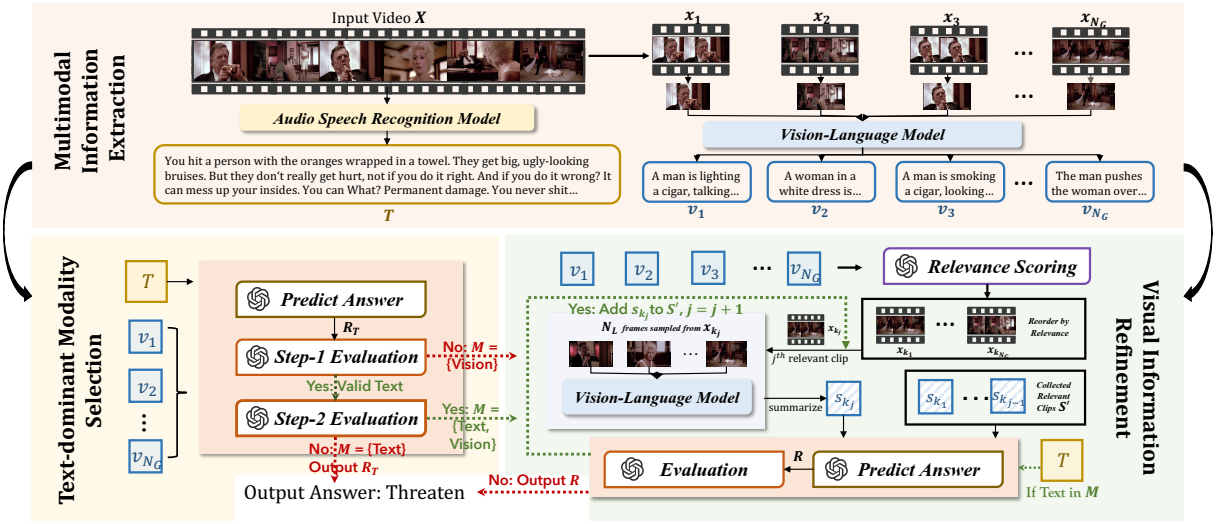

Figure 2: An overview of CFMiS. Here, T refers to textual dialogue, x_i denotes the i_{th} clip and v_i is the corresponding caption, S is the caption summary of the whole video, M is the modality selection result, s_{k_j} is the summary of the j_{th} relevant clip and S' represents the collection of relevant clip summaries, R_T and R are the VSSR prediction result where the former is text-based.

Wang et al., 2023a), mamba (Li et al., 2024), to capture video spatio-temporal clues, but struggle to distinguish similar conversation styles with merely visual data; 2) Multimodal-based methods, which leverage large-scale representation learning (Sun et al., 2022; Argaw et al., 2023; Singh et al., 2024; Wang et al., 2025b), external knowledge (Zhang et al., 2023; Huang et al., 2025; Singh et al., 2025), or pre-trained model finetuning (He et al., 2024; Zhang et al., 2025; Man et al., 2025) to enhance differentiation.

Despite progress, VSSR remains challenging due to its fine-grained nature and limited annotated data. Recent works (Bhattacharya et al., 2023; Lin et al., 2024) explore leveraging LLM with a training-free manner for VSSR. But even with summarization to reduce duplication (Bhattacharya et al., 2023), misleading visual information still persists, overshadowing style-relevant critical clues during multimodal fusion. In contrast, our CFMiS identifies and integrates valuable multimodal information relevant to VSSR, effectively mitigating the negative impcts caused by information redundancy.

Large Language Models for Video Understanding. Motivated by the revolutionized contribution of LLM to natural language processing (Qin et al., 2023; Sun et al., 2024b), via converting visual data into language, recent works explore utilizing LLM to address video understanding tasks with a training-free pipeline (Zhang et al., 2024a; Fan et al., 2024; Zanella et al., 2024).

Information redundancy is an inherent challenge for video understanding. To solve this issue, LifelongMemory(Wang et al., 2023b) eliminates query-irrelevant visual captions based on representation similarity. VideoAgent(Wang et al., 2024) iteratively extracts useful frames via LLM-based evaluation. VideoTree(Wang et al., 2025c) incorporates multi-granularity visual information into a tree-based structure. Unlike these works, CFMiS highlights the dominant position of text for VSSR, taking textual dialogue as reference for valuable information selection, which is specifically designed for VSSR to enhance conversation understanding.

3 Method

The overall framework of CFMiS is illustrated in Figure 2, which outputs the speaking style recognition result of the input conversation video. There are three main components in CFMiS: 1) multimodal information extraction module where textual and visual information of each video are obtained in language form using multimodal perception models in this module; 2) text-dominant modality selection module that firstly assesses if textual information is valid and then evaluates if visual information is additionally required for VSSR; and 3) visual information refinement module where visual contents are graded by the relevance to VSSR and high-relevant visual information are iteratively gathered until LLM generating an absolutely confident VSSR answer.

3.1 Multimodal Information Extraction

Due to some subtitle missing in Youtube website, to obtain complete textual dialogue T, we employ an automatic speech recognition (ASR) model on the audio of each video X, which is extracted via FFmpeg tools as follows:

$$T = \mathcal{F}_{ASR}\left(\mathcal{F}_{FFmpeg}\left(X\right); \theta^{ASR}\right), \quad (1)$$

where $\mathcal{F}_{ASR}$ and θ^{ASR} refer to the ASR model and its parameters, $\mathcal{F}_{FFmpeg}$ represents FFmpeg command tools (Bellard, 2000).

The input video X is firstly segmented into N_G clips in average to obtain the clip sequence $\tilde{X} = \{x_1, x_2, ..., x_{N_G}\}$. The middle frame $f_i, i \in \{1, 2, ..., N_G\}$, of each clip x_i is gathered to represent the video, which is subsequently fed to VLM to generate caption v_i, converting visual data into language:

$$v_i = \mathcal{F}_{VLM}\left(\left[Prompt_C; f_i\right]; \theta^{VLM}\right), \quad (2)$$

where $\mathcal{F}_{VLM}$ and θ^{VLM} refer to the VLM model and its parameters, $Prompt_C$ represents the exclusive prompt that instructs the VLM model to generate captions for video frames, which is "*Caption the image*".

3.2 Text-dominant Modality Selection (TdMS)

From textual dialogue T and visual captions $V = \{v_1, v_2, ..., v_{N_G}\}$, CFMiS firstly finds out valuable information for VSSR by a coarse manner. It proposes a text-dominant modality selection module to evaluate which modality is required for VSSR originating from the text-based prediction.

Text-based Answer Prediction. To generate the text-based prediction result, CFMiS formulates VSSR as a multiple-choice problem for LLM understanding. Specifically, given textual dialogue T, CFMiS instructs LLM to choose from multiple speaking style categories $C = \{c_1, c_2, ..., c_{N_C}\}$, N_C is the total category number, for a VSSR-related query Q. Only one choice R_T is allowed to output as the speaking style recognition result.

Step-1 Evaluation. To avoid invalid textual dialogue, the modality selection process is composed by two steps, where the first step is to check the status of textual information. We feed the given information T, query Q, choice C and the text-based chosen result R_T to LLM, instructing LLM to verify if the answer-producing process is valid for the question, mediately confirming the validity of textual dialogue. Response of this evaluation step is enforced to be either "*yes*" or "*no*", where "*yes*" means textual dialogue is valid while "*no*" means visual captions V will be the only available information source to generate VSSR result.

Step-2 Evaluation. When text modality is retained after the first-step evaluation, the second step activates to evaluate if visual information can make additional contribution. In case of similar frames, CFMiS leverages LLM to summarize V to obtain vision summary S, which represents the whole video without duplicate visual information. We then apply LLM to evaluate if S can make difference for the existing text-based answer-producing process, which is composed by textual dialogue T, VSSR-related query Q, choice C and the text-based chosen result R_T.

LLM output of the second-step evaluation step is also restricted to "*yes*" and "*no*", where "*yes*" response leads to textual and visual information integration for VSSR while "*no*" indicates useless visual information and R_T will become the ultimate prediction result. Coordination of the aforementioned two steps achieves a dynamic modality selection result M for VSSR, which is $\{M_V\}$, $\{M_T\}$ or $\{M_V, M_T\}$, where M_V and M_T refer to vision and text modality, respectively.

Algorithm 1 Iterative visual information refinement

Require: Reordered video clip set $\tilde{X}'$, clip sample number N_L, caption model $\mathcal{F}_{VLM}$, LLM model $\mathcal{F}_{LLM}$, max iteration number N_M, modality selection result M, textual dialogue T, VSSR-related query Q

Let $j \leftarrow 1$, $S' \leftarrow$ set()
while $j \leq N_M$ **do**
  $x_{k_j} \leftarrow \tilde{X}'[j]$.
  $V^{k_j} \leftarrow$ Caption$(\mathcal{F}_{VLM}, $Sample$(x_{k_j}, N_L))$
  $s_{k_j} \leftarrow$ Summarize$(\mathcal{F}_{LLM}, V^{k_j})$
  if $j == 1$ **then**
    $S' \leftarrow S'$.append(s_{k_j})
    continue
  else
    if $M_T \in M$ **then**
      $I =$ Integrate(S', T)
    else
      $I =$ Integrate(S')
    end if
    $R =$ LLMReason(F_{LLM}, I, Q)
    $E =$ LLMEvaluate$(F_{LLM}, I, Q, R, s_{k_j})$
    if *E is positive* **then**
      $S' \leftarrow S'$.append(s_{k_j}), $j \leftarrow j + 1$
    else
      break
    end if
  end if
end while
return R

3.3 Visual Information Refinement (VIR)

If vision is included in the selected modalities M, the visual information refinement module of

CFMiS will be active to find out valuable visual clues for VSSR by a fine manner.

Relevance Scoring. Humans answer questions following a global-to-local process (Wang et al., 2024, 2025c): quickly overview the whole context and then closely observe the question-relevant parts. Inspired by this, we firstly obtain the overview of the whole video by feeding all clip captions $V = \{v_1, v_2, ..., v_{N_G}\}$ and the VSSR-related query Q into LLM, outputting a set of relevance scores $H = \{h_1, h_2, ..., h_{N_G}\}$. We rank all clips based on the relevance score to obtain a reordered clip set $\tilde{X}' = \{x_{k_1}, x_{k_2}, ..., x_{k_{N_G}}\}$, where k_j represents the index of the clip whose relevance score ranks j_{th} in the original clip sequence $\tilde{X}$.

Iterative Visual Information Collection. Since each video carries different quantity of VSSR-relevant clips (Zhang et al., 2025), as shown in Algorithm 1, CFMiS designs an iterative solution to dynamically collect relevant clips: 1) to obtain a close observation on critical clip, we firstly evenly sample N_L frames from current clip x_{k_j} to convert them into language-form captions $V^{k_j} = \{v_1^{k_j}, ..., v_{N_L}^{k_j}\}$ via Equation (2) and we then feed V^{k_j} to LLM to remove duplication via generating clip vision summary s_{k_j}; 2) summaries of all previously collected clips $S' = \{s_{k_1}, s_{k_2}, ..., s_{k_{j-1}}\}$ are chronologically integrated as input to LLM to answer the VSSR-related query Q, obtaining a vision-based prediction result R; and 3) based on R, we leverage LLM to evaluate if the clip summary s_{k_j} can make additional contribution. A positive response means x_{k_j} is a relevant clip to be collected in S' and step 1-3 will be repeated for $x_{k_{j+1}}$ to obtain a more reliable answer. A negative response or achieving max iteration number N_M can prevent the iterative process and the ultimate R will be output as an absolutely confident answer.

Discussion about VIR Efficiency. Actually, according to the modality selection result M, since textual dialogue takes the dominant role in VSSR, small quantity of samples require visual information to participate in visual information refinement (VIR). Furthermore, among these samples, most of them can obtain enough visual details within only one iteration. Hence, even though VIR is an iterative module, it will not substantially increase the overall inference budget of CFMiS, which is also demonstrated in Experiment 4.3.

Method	Acc	F1	P	R	WF1	WP
Training-based						
ObjTrans(Wu and Krahenbuhl, 2021)	40.3	35.7	36.2	36.4	39.1	38.7
ViS4mer (Islam and Bertasius, 2022)	38.3	32.9	35.3	34.3	36.3	37.2
LF-VILA (Sun et al., 2022)	40.3	31.9	31.1	34.1	37.6	36.6
S5 (Wang et al., 2023a)	42.1	-	-	-	-	-
M2S (Chen et al., 2023)	42.2	-	-	-	-	-
LMP (Argaw et al., 2023)	44.4	-	-	-	-	-
MMSF (Zhang et al., 2023)	50.2	45.0	48.0	44.5	49.1	49.5
MA-LLM (He et al., 2024)	41.2	36.4	40.4	38.1	39.0	42.4
LSSD (Singh et al., 2024)	50.8	-	-	-	-	-
VideoMamba (Li et al., 2024)	38.3	27.6	29.8	30.0	34.4	33.1
B-LLaVA (Singh et al., 2025)	41.0	-	-	-	-	-
TNvE (Zhang et al., 2025)	**56.7**	**51.7**	**56.8**	**53.3**	**54.8**	**57.9**
Training-free						
LLoVi (Zhang et al., 2024a)	32.8	21.2	29.8	22.6	28.6	37.7
VideoTree (Wang et al., 2025c)	33.8	28.1	38.0	28.3	30.9	36.1
4096T (Bhattacharya et al., 2023)	40.3	36.6	46.5	34.6	40.1	46.7
MovieSeq (Lin et al., 2024)	43.3	39.7	45.7	37.8	47.2	57.3
CFMiS (Ours)	**53.7**	**43.0**	**46.9**	**45.5**	**50.5**	**57.4**

Table 1: Performance comparison results of our CFMiS *vs.* different state-of-the-art methods on LVU-VSSR dataset. Here, best results of training-based and training-free methods are in bold, respectively.

4 Experiments

4.1 Datasets and Experimental Settings

Datasets. Consistent with previous works (Zhang et al., 2025), we conduct experiments on LVU-VSSR and LVU-VSRR datasets (Wu and Krahenbuhl, 2021). LVU-VSSR is a widely-used VSSR dataset, comprising approximately 1,000 conversation videos, each of which ranges from one to three minutes. There are five distinct speaking style categories and each video is labeled with one speaking style. LVU-VSRR is a video social relationship recognition dataset. Since social relationship among humans can also be revealed by key behaviors, LVU-VSRR is leveraged to validate the generalization of CFMiS. LVU-VSRR contains approximately 200 human-centered videos, each of which lasts one to three minutes. It involves four social relationship categories and each video relates to one relationship label.

Evaluation Metrics. We utilize six standard classification metrics to evaluate VSSR performance: Top-1 Accuracy (Acc), Macro F1-score (F1), Macro Precision (P), Macro Recall (R), Weighted F1-score (WF1) and Weighted Precision (WP). For all these metrics, higher value signify superior performance.

Implementation Details. The ASR model is Whisper (Radford et al., 2023) and the VLM model is CogAgent (Hong et al., 2024). The LLM model is GPT-4o Mini (OpenAI, 2024a). The video clip number N_G is 8 and the clip frame number N_L is 4. Query Q related to task is

Method	Acc	F1	P	R	WF1	WP
Training-based						
ObjTrans(Wu and Krahenbuhl, 2021)	54.8	28.4	35.0	36.1	41.3	43.1
ViS4mer (Islam and Bertasius, 2022)	57.1	40.3	36.4	45.2	51.4	46.7
LF-VILA (Sun et al., 2022)	57.1	33.2	51.7	38.9	45.3	57.4
S5 (Wang et al., 2023a)	67.1	-	-	-	-	-
M2S (Chen et al., 2023)	71.2	-	-	-	-	-
LMP (Argaw et al., 2023)	69.4	-	-	-	-	-
MA-LLM (He et al., 2024)	57.9	34.8	47.6	40.2	47.9	59.2
LSSD (Singh et al., 2024)	61.0	-	-	-	-	-
VideoMamba (Li et al., 2024)	57.1	35.1	34.3	40.2	47.6	43.7
B-LLaVA (Singh et al., 2025)	64.2	-	-	-	-	-
TNvE (Zhang et al., 2025)	**71.4**	**61.5**	**80.0**	**59.6**	**68.1**	**75.7**
Training-free						
LLoVi (Zhang et al., 2024a)	54.8	35.2	36.8	34.9	54.5	55.7
VideoTree (Wang et al., 2025c)	59.5	42.6	45.4	43.1	62.2	68.3
4096T (Bhattacharya et al., 2023)	66.7	49.0	53.5	48.5	67.3	72.3
MovieSeq (Lin et al., 2024)	57.1	36.3	39.3	34.2	62.6	70.5
CFMiS (Ours)	**71.4**	**54.3**	**63.3**	**48.7**	**77.2**	**86.5**

Table 2: Performance comparison results of our CFMiS *vs.* different state-of-the-art methods on LVU-VSRR dataset.

Method	Inf Time (s/sample)	Inf Expense ($/sample)	Frames	Acc
LLoVi (Zhang et al., 2024a)	33.1	5.8e-2	68	32.8
4096T (Bhattacharya et al., 2023)	18.7	1.6e-3	14	40.3
CFMiS (Ours)	26.4	9.0e-4	12	**53.7**
CFMiS w/o Sum	**8.6**	**4.0e-4**	**9**	51.7

Table 3: LLM inference time and expense comparison results of our CFMiS *vs.* different training-free methods on LVU-VSSR dataset. Here, Sum denotes caption summarization in CFMiS.

{"*What is the speaking style of the conversation in this video?*", "*What is the relationship between the characters in this video?*"} for LVU-VSSR and LVU-VSRR, respectively. The max iteration number N_M is 5. LLM prompt details can be obtained in appendix A.2.

4.2 Comparison with State-of-the-Arts

As shown in Table 1 and 2, we compare CFMiS with state-of-the-art baselines involving both training-based and training-free methods (Section A.3.1). It can be observed that CFMiS outperforms all training-free approaches for all metrics, which demonstrates the effectiveness of CFMiS in enhancing LLM reasoning for VSSR. Even compared to training-based methods, CFMiS is superior to most of them, except for TNvE, which is supervised trained with more modalities, *e.g.,* audio. It is noteworthy that the LLM model CFMiS utilizes is GPT-4o Mini, the same as MovieSeq (Lin et al., 2024). Moreover, as mentioned in Table 7, when replacing the LLM model of CFMiS to GPT-3.5, which is the LLM backbone of 4096Tokens (Bhattacharya et al., 2023), CFMiS still performs better. These results prove that the primary advantage of CFMiS lies

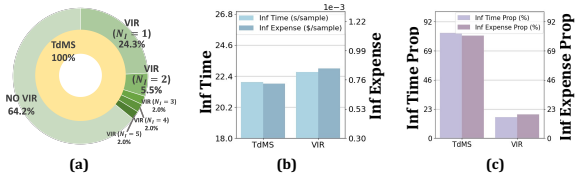

Figure 3: Involved sample distribution, LLM inference time, expense and corresponding proportion in total CFMiS of TdMS *vs.* VIR. Here, N_I denotes the iteration number of VIR.

Model	Mod.	TdMS	VIR	Acc	F1	R	WF1
Modality Ablation							
Base$_V$	V	✗	✗	33.8	27.2	29.7	27.4
Base$_T$	T	✗	✗	50.3	40.8	42.3	48.3
Base$_M$	V, T	✗	✗	49.8	40.1	42.4	46.4
Module Ablation							
CFMiS	V, T	✓	✓	**53.7**	**43.0**	**45.5**	**50.5**
CFMiS w/o TdMS	V, T	✗	✓	52.2	41.3	43.2	49.5
CFMiS w/o VIR	V, T	✓	✗	51.2	41.1	43.3	48.4
Other Components Ablation							
CFMiS w/o Sum	V, T	✓	✓	51.7	41.6	43.7	48.8
CFMiS w/o COT	V, T	✓	✓	52.2	41.7	43.8	49.5
CFMiS w/ DC	V, T	✓	✓	51.7	41.6	43.4	48.8

Table 4: Ablation results of different CFMiS variants on LVU-VSSR dataset. Here, V and T refer to vision and text modalities, Sum and DC denote Summarization and Detect Clip.

in the design of dynamic multimodal information selection, not leveraging the advanced LLM model for VSSR.

4.3 Efficiency Analysis

To prove the efficiency of CFMiS, we compare the LLM inference time and expense of CFMiS with other training-free methods. In Table 3, the inference time of CFMiS is moderate but its expense is the lowest and accuracy is the best. In addition, due to the information selection mechanism, CFMiS involves the least visual frames which means the lowest caption cost. Actually, CFMiS computation budget mainly comes from the caption summarization outputs, which can be significantly reduced without summarization and still keeps a high accuracy.

For a deeper analysis, although VIR is a dynamic iterate module, as Figure 3(a) shows, small quantity of samples require visual information to take part in VIR and only 11.5% iterate for over one round. This is because for VSSR, subtitle text and tiny key visual clues make main contribution. Only few hard samples require multiple VIR iterations to collect more available visual clues. Thus, as shown in 3(b), the iterative VIR inference time and expense are similar with those of the non-iterative TdMS module. And in Figure 3(c), VIR only

Text-dominant Modality Selection			
Procedure	Step-1 Eval_V	Step-2 $\text{Eval}_\text{V,T}$	Step-2 Eval_T
Precision	66.7	83.3	78.5
Visual Information Refinement			
Procedure	HitRate@1	HitRate@3	HitRate@5
Relevance Scoring	55.2	79.1	90.3

Table 5: Quantitative analysis results about the LLM response quality of CFMiS on LVU-VSSR dataset. Here, Step-1 Eval_V, Step-2 $\text{Eval}_\text{V,T}$ and Step-2 Eval_T refer to modality selection results which are only vision, vision combines text and only text, respectively.

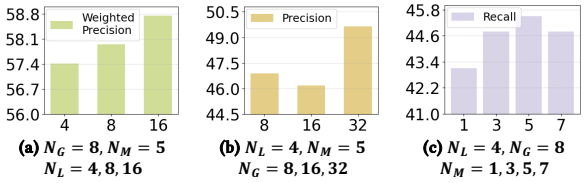

Figure 4: Ablation results of CFMiS with different sampling rates N_G, N_L and maximum iterative number N_M on LVU-VSSR dataset.

accounts for 16.7% / 18.9% of the total CFMiS inference time/expense, which is a relatively low-cost module in practice.

4.4 Ablation Study

Modality, Module and Other CFMiS Components. We conduct various ablation studies to demonstrate the effectiveness of each component in CFMiS for VSSR. The comparison results shown in Table 4 lead to the following observations.

The base models are implemented by directly inputting the textual dialogue or visual summary of the whole video to LLM to answer the VSSR-related query. It can be observed that applying textual information outperforms visual information by a large margin, which proves the dominant role of text for VSSR. Moreover, equally integrating all textual and visual information leads to a performance drop, caused by redundancy and noise in multimodal information.

When taking out any one of the TdMS module and the VIR module, the VSSR performance declines but still outperforms the base multimodal model, which verifies that both of these two modules take effect for multimodal information selection and combining them together can maximize the VSSR outcome.

For other ablation components in CFMiS: 1) when inputting original frame captions to LLM without any summarization, the VSSR performance obviously decreases, which proves that eliminating similar frames is necessary for VSSR; 2) by including "*Think step by step*" in LLM prompts and instructing LLM to output the reasoning process, CFMiS incorporates chain of thought (COT) to avoid LLM hallucination (Wei et al., 2022), achieving positive influence; and 3) we utilize TransNetV2 (Souček and Lokoč, 2020) to divide video into multiple clips by scene detection, as an alternate of the average segmentation in CFMiS. However, the VSSR performance declines which we think is because conversation videos exhibit visually consistent for most of the time. Hence, scene-based clips are too long where the sparse critical visual clues tend to be overlooked.

Model	#param	Acc	F1	R	WF1
LaViLa (Zhao et al., 2023)	N/A	49.3	39.8	41.6	46.6
LLaVA1.5-7B (Liu et al., 2024)	7B	50.8	40.2	42.2	47.3
LLaVA1.5-13B (Liu et al., 2024)	13B	51.7	41.6	43.3	48.7
Qwen3VL-8B (Bai et al., 2023)	8B	52.7	42.4	44.1	49.7
Qwen3VL-32B (Bai et al., 2023)	32B	50.8	40.9	42.6	48.4
BLIP2-T5-XL (Li et al., 2023)	N/A	50.8	40.7	42.6	47.8
BLIP2-T5-XXL (Li et al., 2023)	N/A	51.7	41.5	43.6	48.8
CogAgent (Hong et al., 2024)	18B	**53.7**	**43.0**	**45.5**	**50.5**
Raw Image	-	52.2	42.2	44.5	49.4

Table 6: Ablation results of CFMiS leveraging different VLM models to generate visual captions on LVU-VSSR dataset. Here, 1B denotes one billion parameters.

LLM Response Quality. Since CFMiS is a training-free framework relying on LLM. To verify the robustness of CFMiS, we quantitatively evaluate the LLM response quality of each procedure in CFMiS. As for the LLM responses of text-dominant modality selection module, we directly leverage LLM to predict speaking style based on contents of selected modality and positive prediction result means the modality selection result is correct. As for the LLM response of relevance scoring in visual refinement module, we predict speaking style based on high-relevant clips and positive prediction result means correct scoring. Encouraging evaluation results in Table 5 prove the reliability of LLM responses, further demonstrating the robustness of CFMiS.

Sampling Rates and Max Iteration Number. Results of ablation study on different N_L, N_G and N_M are respectively shown in Figure 4(a)-(c), leading to the following observations: 1) VSSR performance increases with higher N_L, as sampling more frames from relevant clips can include more useful details for VSSR; 2) with the highest N_G, VSSR performance achieves the best, because the most fine-cut clips are most likely to involve critical visual clues that stand out in clip relevance scoring; and 3) VSSR performance keeps growing up until

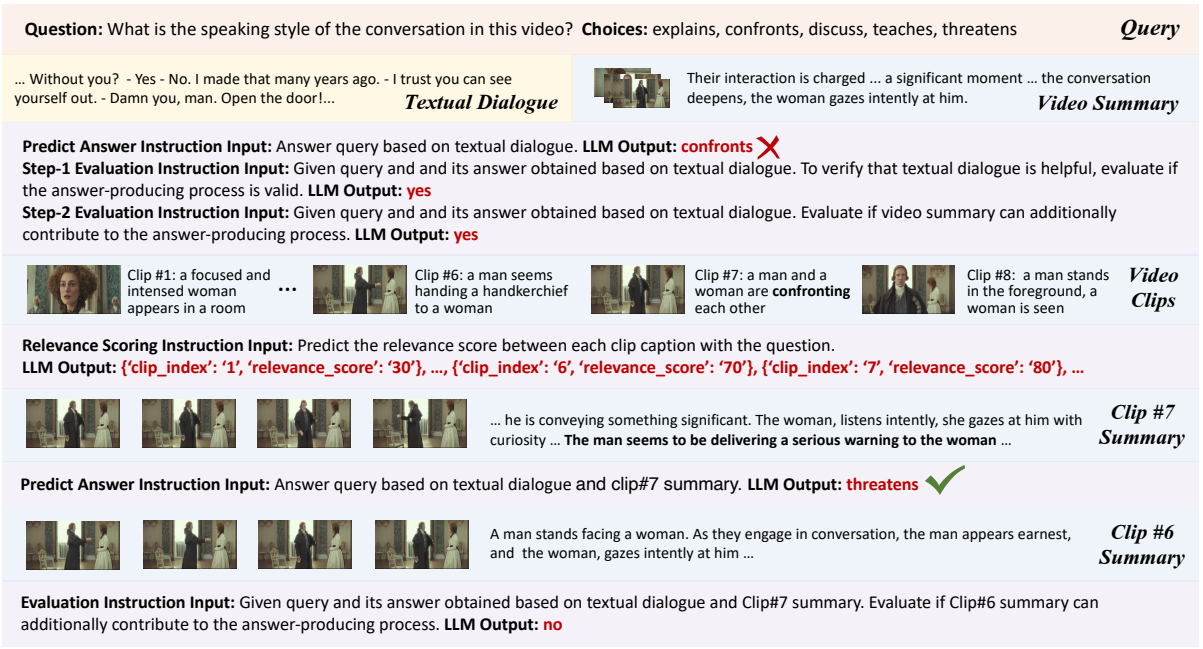

Figure 5: An example of the answer producing process of CFMiS for VSSR on LVU-VSSR dataset.

Model	Type	#param	Acc	F1	WF1
Open-source LLM					
LLaMA3-8B (AI@Meta, 2024)	Chat	8B	44.8	38.1	43.8
LLaMA3-70B (AI@Meta, 2024)	Chat	70B	44.3	35.2	41.6
Qwen3-32B (Yang et al., 2025)	Chat	32B	49.3	41.6	43.7
DeepSeek-R1 (DeepSeek-AI, 2025)	Think	671B	42.8	43.8	41.1
Proprietary LLM					
GPT-3.5 (Brown et al., 2020)	Chat	N/A	46.8	35.0	40.5
GPT-4o Mini (OpenAI, 2024a)	Chat	N/A	53.7	43.0	50.5
GPT-4.1 (OpenAI, 2024b)	Chat	N/A	53.2	45.6	54.4
GPT-4o (OpenAI, 2024c)	Chat	N/A	**55.2**	**46.1**	**56.0**

Table 7: Ablation results of CFMiS leveraging different LLM for VSSR on LVU-VSSR dataset. Here, 1B denotes one billion parameters.

$N_M = 5$, since the number of relevant clips is limited, interference is more likely to occur with more irrelevant clips involved.

Caption Models. To assess the impact of caption quantity produced by various VLMs on VSSR performance, we examine multiple outstanding VLMs along with different model scales, as shown in Table 6. The experimental results verify that CogAgent achieves the best. In addition, we directly feed frame images, without any language-form conversion, to GPT-4o Mini. Declined performance in the last line of Table 6 demonstrate that converting visual data into language is better for large models to understand and reason for VSSR.

LLMs. To evaluate how different LLMs influence the VSSR performance of CFMiS, we experiment on various LLMs, including open-source and proprietary models with different parameter quantities. The experimental results are shown in Table 7. It can be observed that most proprietary models outperform open-source ones except for the outdated GPT-3.5. GPT-4o Mini is a cost-effective LLM model for CFMiS with significantly fewer parameters, applied for most experiments. As the enhanced version of GPT-4o Mini, GPT-4o achieves an evident improvement, revealing the limitless potential of CFMiS with the development of LLMs.

4.5 Visualization

Figure 5 visualizes the answer producing process of CFMiS. Firstly, CFMiS selects both text and vision as VSSR-required modalities via confirming the validity of textual dialogue and video summary. Subsequently, CFMiS generates the relevance scores of all visual clips and the seventh clip is highlighted due to the confrontational atmosphere. To take a close view, CFMiS samples more frames from the seventh clip, obtaining critical visual clues "a serious warning" and achieving a correct prediction "*Threaten*". Since the sub-relevant clip cannot provide distinct evidence for VSSR, "*Threaten*" is considered as the ultimate output. The above process qualitatively proves the effectiveness of CFMiS for VSSR, especially for selecting valuable information. More visualization cases can be obtained in appendix A.3.2.

5 Conclusion

In this paper, we proposed CFMiS, a coarse-to-fine multimodal information selection framework, to dynamically obtain valuable information for VSSR with large language model. It involves a text-dominant modality selection module and a visual information refinement module, where the former resolves which modality to utilize while the latter further determines which information to adopt from the selected modalities. Extensive experiments and ablation studies were conducted to demonstrate the effectiveness of CFMiS.

Limitations

Although CFMiS achieves advanced performance in VSSR, it also exhibits some specific constraints. First, summarization and COT outputs cause certain LLM inference latency and expense. However, this can be alleviated with output restriction or applying recent omni models (Xu et al., 2025) with raw data as input to avoid summarization. Second, LLM hallucinations and instability may cause inaccurate relevance scores or inconsistent results, undermining information filtering of CFMiS and destroying VSSR accuracy. While CoT prompting and verification mitigate this, full resolution remains so challenging that future work should focus on enhancing LLM output credibility, alongside advancing LLM alignment and designing more validation mechanisms to enhance reliability.

Acknowledgements

This work is supported by the National Science Foundation of China (92582103, 62072232), the Industrialization and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM118), and the Collaborative Innovation Center of Novel Software Technology.

References

AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Deepali Aneja, Rens Hoegen, Daniel McDuff, and Mary Czerwinski. 2021. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 1–10.

Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, and Fabian Caba Heilbron. 2023. Long-range multimodal pretraining for movie understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13392–13403.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. In *arXiv preprint arXiv:2308.12966*.

Fabrice Bellard. 2000. Ffmpeg. http://ffmpeg.org.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize videos to understand them in zero shot. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9822–9839.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 1877–1901.

Shixing Chen, Chun-Hao Liu, Xiang Hao, Xiaohan Nie, Maxim Arap, and Raffay Hamid. 2023. Movies2scenes: Using movie metadata to learn scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6535–6544.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. In *Nature*, volume 645, pages 633–638.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 75–92.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. In *arXiv preprint arXiv:2306.08640*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14281–14290.

Wei-Lun Huang, Shintami Chusnul Hidayati, and Tse-Yu Pan. 2025. Movie retrieval systems using genre-guided multimodal learning techniques. In *Proceedings of the International Conference on Multimedia Modeling*, pages 158–164.

Md Mohaiminul Islam and Gedas Bertasius. 2022. Long movie clip classification with state-space video models. In *Proceedings of the European Conference on Computer Vision*, pages 87–104.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. Videomamba: State space model for efficient video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 237–255.

Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. 2024. Learning video context as interleaved multimodal sequences. In *Proceedings of the European Conference on Computer Vision*, pages 375–396.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. 2025. Adacmˆ 2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8544.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Introducing gpt-4.1 in the api.

OpenAI. 2024c. Introducing gpt-4o and more tools to chatgpt free users.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, pages 28492–28518.

Sepehr Sameni, Simon Jenni, and Paolo Favaro. 2023. Spatio-temporal crop aggregation for video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5664–5674.

Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Yan Song, Zechao Li, and Liyan Zhang. 2017. Concurrence-aware long short-term sub-memories for person-person action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Nikhil Singh, Chih-Wei Wu, Iroro Orife, and Mahdi Kalayeh. 2024. Looking similar sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26907–26918.

Somesh Singh, Harini SI, Yaman K Singla, Veeky Baths, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. 2025. Teaching human behavior improves content understanding abilities of llms. In *Proceedings of the International Conference on Learning Representations*.

Tomáš Souček and Jakub Lokoč. 2020. Transnet v2: An effective deep network architecture for fast shot transition detection. In *arXiv preprint arXiv:2008.04838*.

Ombretta Strafforello, Klamer Schutte, and Jan Van Gemert. 2023. Are current long-term video understanding datasets long-term? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2967–2976.

Kaili Sun, Zhiwen Xie, Mang Ye, and Huyin Zhang. 2024a. Contextual augmented global contrast for multimodal intent recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26963–26973.

Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2024b. PEARL: Prompting large language models to plan and execute actions over long documents. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 469–486.

Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. 2022. Long-form video-language pre-training with multimodal temporal contrastive learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 38032–38045.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. 2023a. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6387–6397.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2025a. Gpt-ner: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024. Videoagent: Long-form video understanding with large language model as agent. In *Proceedings of the European Conference on Computer Vision*, pages 58—-76.

Yicheng Wang, Zhikang Zhang, Jue Wang, David Fan, Zhenlin Xu, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. 2025b. Gexia: Granularity expansion and iterative approximation for scalable multi-grained video-language learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4725–4735.

Ying Wang, Yanlai Yang, and Mengye Ren. 2023b. Lifelongmemory: Leveraging llms for answering queries in long-form egocentric videos. In *arXiv preprint arXiv:2312.05269*.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025c. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3283.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 24824–24837.

Wen-Li Wei, Chung-Hsien Wu, Jen-Chun Lin, and Han Li. 2013. Interaction style detection based on fused cross-correlation model in spoken conversation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8495–8499.

Wen-Li Wei, Chung-Hsien Wu, Jen-Chun Lin, and Han Li. 2014. Exploiting psychological factors for interaction style recognition in spoken conversation. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 22, pages 659–671.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894.

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. 2022. Hierarchical self-supervised representation learning for movie understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9727–9736.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. In *arXiv preprint arXiv:2509.17765*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. In *arXiv preprint arXiv:2505.09388*.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536.

Beibei Zhang, Yaqun Fang, Fan Yu, Jia Bei, and Tongwei Ren. 2023. Mmsf: A multimodal sentiment-fused method to recognize video speaking style. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pages 289–297.

Beibei Zhang, Tongwei Ren, and Gangshan Wu. 2025. Text-guided nonverbal enhancement based on modality-invariant and-specific representations for video speaking style recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22354–22362.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple llm framework for long-range video question-answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the ACM International Conference on Multimedia*, pages 1688—-1697.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2024b. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *Proceedings of the European Conference on Computer Vision*, pages 124–142.

A Appendix

In this supplementary material, we provide the following: 1) related work about the difference and value of VSSR; 2) method details of CFMiS, including all LLM prompt templates; and 3) additional experiments, which are diverse visualization cases, analysis on the audio modality and the task-specificity of CFMiS.

A.1 Related Work

Since some readers may confuse about the difference and value of the VSSR task, we will clarify the significance of VSSR in this section.

Difference of VSSR to Video Intent Detection. VSSR aims to identify the conversation styles among characters in videos which takes video as input and conversation category as output. Different from Intent detection (Zhang et al., 2022; Sun et al., 2024a), which aims to recognize the intention of a single person when he/she interacts with others. VSSR focus on recognizing the conversation interaction type involved multiple people, such as "Discuss" and "Confront". Thus, intent detection mostly deal with short videos (few seconds) which just involves single person shot while VSSR process long videos (100 seconds) with multiple shots involving different people. Moreover, intent recognition also include action intents such as "Leave" (Zhang et al., 2022) while VSSR only study on conversation type. Thus, VSSR methods (Zhang et al., 2025) take textual dialogue as the primary role while other modalities tend to be redundant. However, intent detection methods (Sun et al., 2024a) regard visual and acoustic information equally important, aiming to capture emotions from visual expressions and acoustic tones.

Value of VSSR. According to the statistics (Zhang et al., 2023), each person on average takes part in conversation of 14,878 words each day, lasting about an hour and fifteen minutes, which proves that conversation accounts for a large proportion in human daily lives. For more comprehensive conversation understanding, (Wu and Krahenbuhl, 2021) proposed video speaking style recognition task (VSSR), whereas previous works concentrate on analyzing speaking styles in speeches. Different from conventional action recognition (Shu et al., 2017), VSSR offers a distinctive and essential perspective on human interactions within conversations. It serves a significant role in numerous applications, including video understanding applications like video tagging and other specific scenarios like conversation head generation. Consequently, VSSR has garnered more and more interest in recent years (Zhang et al., 2023, 2025).

A.2 LLM Prompts

In this section, we will provide all specific LLM prompt templates used in CFMiS.

The answer prediction prompt in text-dominant modality selection module. Here, T refers to textual dialogue, Q is the task-related query and C is the category set.

> *You are presented with the subtitle of a video. The ultimate goal is to answer a question related to this video. Please think step-by-step and choose the most relevant answer. Note that only one answer is returned for the question.Subtitle: T; Question: Q; Choices: C. Please output in JSON format {"final_answer": "xxx", "thinking": "xxx"}.*

The first-step evaluation prompt in text-dominant modality selection module. Here, R_T is the text-based prediction result.

> *You are presented with a question and its answer obtained based on the given information. To verify that the given information is helpful, please evaluate if the answer-producing process is valid for the question. Question: Q; Choices: C; Given information: T; Answer-producing: R_T. Think step by step and output yes or no, where yes indicates the answer-producing is valid for the question based on helpful information, and no indicates invalid. Please output in JSON format {"valid": "xxx", "thinking": "xxx"}.*

The summarization prompt in text-dominant modality selection module. Here, V is the collection of all clip captions.

> *You are presented with some frame captions of a video in temporal order. Please write a coherent and narrative summary based on these captions. Use only the information provided and make sure the summarization feels like a continuous narrative. Do not include any details not mentioned in the prompt. Captions: V.*

The second-step evaluation prompt in text-dominant modality selection module. Here, S is the video caption summary.

> *You are presented with a question and its answer obtained based on the given information. Please evaluate if new information can additionally contribute to the answer-producing process for the question. Question: Q; Choices: C; Given information: T; Answer-producing: R_T; New information: S. Think step by step and output yes or no, where yes indicates new information can make additional contribution to the answer-producing process except for the given information, and no indicates there is no need for new information since given information is sufficient to obtain a high-confident answer-producing for the question. Please output in JSON format {"contribute": "xxx", "thinking": "xxx"}.*

The relevance scoring prompt in visual information refinement module.

> *You are presented with the textual description of a video, which consists of about some frame captions sampled from the video. The ultimate goal is to answer a question related to this video. Please think step by step and predict the relevance score between each frame caption with the question. Notes that the relevance score ranges from 1 to 100, higher score means more relevant. It is crucial that you imagine the visual scene as vividly as possible to enhance the accuracy of your response. Description: V; Question: Q. Please output in JSON format {"frame_index": "xxx", "relevance_score": "xxx", "thinking": "xxx"}. Note that each frame requires one prediction output and do not output same relevance score.*

The summarization prompt in visual information refinement module is the same as that in text-dominant modality selection module, where the input caption collection is extracted from the sampled frames of a single clip.

The answer prediction prompt in visual information refinement module. Here, S' refers to the summary set of collected relevant clips. If text is included in the modality selection result M, textual dialogue will be involved in this prompt with the form *"Subtitle: T"*.

> *You are presented with the textual description of a video, which consists of about some frame captions sampled from the video. The ultimate goal is to answer a question related to this video. Please think step-by-step and choose the most relevant answer. Note that only one answer is returned for the question. Description: S'; Question: Q; Choices: C. Please output in JSON format {"final_answer": "xxx", "thinking": "xxx"}.*

The evaluation prompt in visual information refinement module is similar to the second-step evaluation prompt in text-dominant modality selection module, where the given information is formed in JSON format *{"Visual_Description": S'}*. When text is included in the modality selection result, *"Subtitle: T"* will be involved in the given information. In addition, provided new information is s_{k_j}, which is the summary of the current clip.

A.3 Additional Experiments

In this section, we will introduce the baselines involved in SOTA comparison with CFMiS. In addition, to make our work more thorough, we provide more visualization cases to show different information selection results of CFMiS and supplement discussions about the audio modality in CFMiS.

A.3.1 Baselines

CFMiS is compared with plentiful methods which can be broadly divided into two categories:

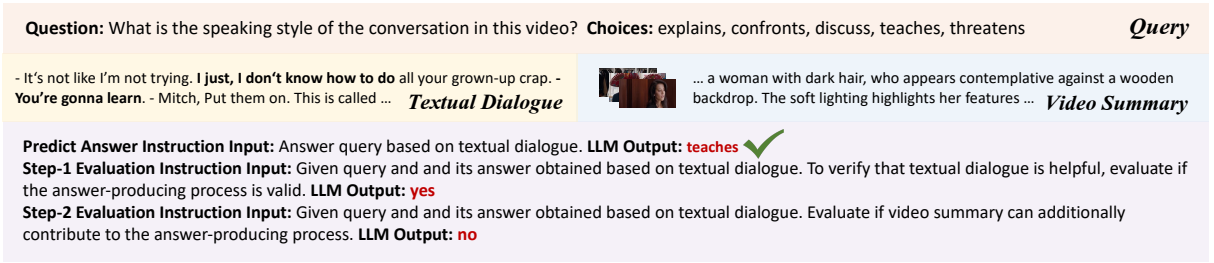

Figure 6: An example of the answer producing process of CFMiS that only selects text modality for VSSR on LVU-VSSR dataset.

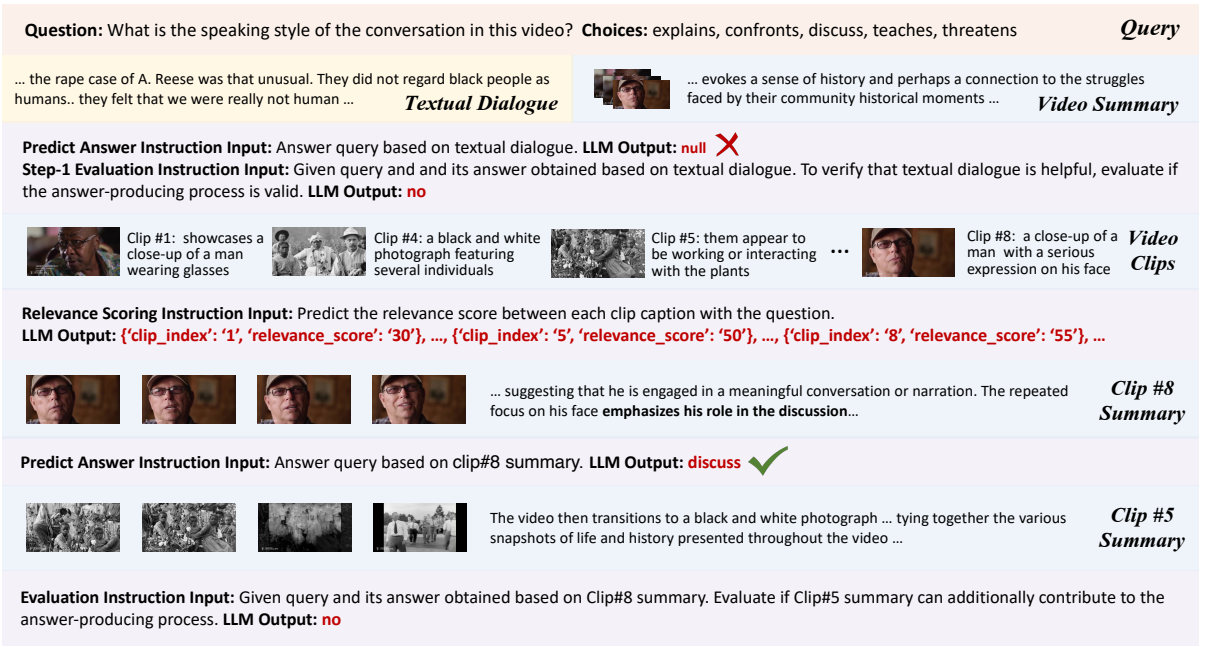

Figure 7: An example of the answer producing process of CFMiS that only selects vision modality for VSSR on LVU-VSSR dataset.

training-based (Wu and Krahenbuhl, 2021; Islam and Bertasius, 2022; Sun et al., 2022; Wang et al., 2023a; Chen et al., 2023; Argaw et al., 2023; Zhang et al., 2023; He et al., 2024; Singh et al., 2024; Li et al., 2024; Singh et al., 2025; Zhang et al., 2025) and training-free methods (Zhang et al., 2024a; Wang et al., 2025c; Bhattacharya et al., 2023; Lin et al., 2024), where the former capitalize on supervised data and the latter explore the potential of LLM for VSSR. Both of these two types of methods focus on either designing sophisticated frameworks to capture robust visual clues or investigating more valuable information from multimodal data sources. Most of these methods are reproduced based on their publicly available codes to obtain evaluation results of various metrics.

A.3.2 Visualization Cases

Figure 6 illustrates that CFMiS only selects textual information for VSSR and outputs the text-based prediction result as the final result. In this case, visual information include few talking scenes, which is hard to recognize speaking styles.

In Figure 7, crime and racial discrimination in dialogue mistake LLM to generate an invalid answer. As a result, CFMiS abandons textual information and focuses on capturing critical visual clues for VSSR. Via scoring the relevance of each clip to VSSR, CFMiS finds out the most relevant clip, the eighth clip, to further obtain details for generating a confident answer.

Except for selecting both text and vision modalities, CFMiS iteratively collects multiple relevant clips in Figure 8. In the first clip, the man and the woman participate in a tense quarrel

Figure 8: An example of the answer producing process of CFMiS that iteratively selects relevant visual clips with two rounds for VSSR on LVU-VSSR dataset.

Method	Dataset	Modality	Acc	WF1
Qwen3-Omni-30B (Xu et al., 2025)	LVU-VSSR	V,A	46.8	45.9
Qwen3-Omni-30B (Xu et al., 2025)	LVU-VSSR	V,T	47.8	47.1
Qwen3-Omni-30B (Xu et al., 2025)	LVU-VSSR	V,T,A	45.3	44.4
CFMiS (Ours)	LVU-VSSR	V,T	**53.7**	**50.5**
Qwen3-Omni-30B (Xu et al., 2025)	LVU-VSRR	V,A,T	59.5	61.9
CFMiS (Ours)	LVU-VSRR	V,T	**71.4**	**77.2**

Table 8: Performance comparison results of our CFMiS *vs.* Qwen3-Omni-30B with different modality combinations on both LVU-VSSR and LVU-VSRR datasets.

while the antagonism between them escalates into physical confrontation in the seventh clip, which jointly corroborates the *"Confront"* style of this video.

A.3.3 Analysis on Audio

In VSSR, most existing works (Zhang et al., 2025; Argaw et al., 2023; Zhang et al., 2023) take visual frames as vision modality, audio signals as audio modality and text subtitle as text modality. Thus, subtitle is considered as text modality even though it is obtained by ASR models due to some subtitle missing in youtube website. Thus, firstly, CFMiS is focused on vision and text.

Secondly, as for audio, researchers normally analyze audio-specific information, such as tone, volume, pitch, which reveal character emotion and sentiment to help recognize speaking styles. Actually we have experimented to include audio in CFMiS by applying various audio understanding models to obtain audio emotion, sentiment and intents. However, when we aggregate these audio-related information with other two modalities (vision and text), the VSSR performance drops. We think this is because of two challenging problems: 1) what type of audio-related information should be extracted? and 2) how to reduce audio redundancy since audio represents steady for most of time in conversaton? We are interested in and will make deeper study in these audio-related problems in our future work. But in CFMiS, we focus on dealing with visual redundancy.

We integrate raw visual frames, audio signals and textual subtitles into Qwen3-Omni-30B (Xu et al., 2025) to solve VSSR. The experimental re-

Model	LVU-VSSR		LVU-VSRR	
	Accuracy	Recall	Accuracy	Recall
Vision-dominant	52.7	44.5	59.5	31.2
Text-dominant	**53.7**	**45.5**	**71.4**	**48.7**

Table 9: Ablation results of taking different modality as the dominator for CFMiS on both LVU-VSSR and LVU-VSRR datasets.

Method	Modality	Accuracy
Training-based		
ObjTrans(Wu and Krahenbuhl, 2021)	V	56.9
LF-VILA (Sun et al., 2022)	V, T	68.0
S5 (Wang et al., 2023a)	V	73.5
LMP (Argaw et al., 2023)	V, A, T	67.8
MA-LLM (He et al., 2024)	V	**80.3**
B-LLaVA (Singh et al., 2025)	V, T	66.4
Training-free		
LLoVi (Zhang et al., 2024a)	V	**84.7**
4096T (Bhattacharya et al., 2023)	V, T	69.4
CFMiS (Ours)	V, T	67.1

Table 10: Performance comparison results of our CFMiS *vs.* different state-of-the-art methods on LVU-Scene dataset.

sults are shown in Table 8, where the performance declines when integrating audio signals, revealing that audio needs further process to take effect.

A.3.4 CFMiS is Task-specific

CFMiS is specific for VSSR since its main motivation is to retain semantic-rich textual dialogue and select few critical visual clues for VSRR. Thus, text modality takes the dominant role in CFMiS, where the modality selection module is a text-dominant module that originates from text-based prediction and integrates visual information only when textual information is insufficient. If we take vision modality as the dominator, the recognition performance drops for both two datasets, as shown in Table 9.

Moreover, for other video understanding tasks, different modalities play different levels of role. For example, vision is as important as text in affect classification since visual expressions closely relate to affect (Zadeh et al., 2017). And for scene recognition task, vision makes much more contribution than text (Lin et al., 2024). We also apply CFMiS for video scene recognition. Experimental results in Table 10 prove that CFMiS performs worse than those regard vision modality as important.