Text-Centric Multimodal Context Modeling for Chinese Patronizing and Condescending Language Detection

Anonymous ACL submission

Abstract

Chinese Patronizing and Condescending Language (CPCL) often appears as sympathy, praise, advice, or moral concern while conveying superiority, pity, or dependence toward vulnerable groups. We study CPCL in short videos as a problem of contextual stance interpretation, with text as the main evidence source. PCLMMPLUS extends PCLMM to 831 videos and adds 103,000 top-level comments, allowing us to examine post-publication social feedback. The model uses ASR transcripts and comments as textual inputs, aligns visual, acoustic, and facial features to the textual state, and combines them with semantic knowledge evidence in a text-centric evidence classifier. On PCLMMPLUS, ablations identify ASR as the central signal, nonverbal cues as auxiliary context, and comments as noise-sensitive feedback. Qwen-based semantic evidence yields the strongest setting, and component diagnostics support a conservative interpretation as sample-level semantic enrichment rather than isolated symbolic graph reasoning. The final model reaches 0.8497 macro-F1 and 0.9296 AUC. These findings suggest that short-video CPCL detection benefits from grounding multimodal evidence in linguistic stance rather than treating all signals as equal video-level cues. Code and dataset materials are available at https://anonymous.4open.science/r/PCLD-747D.

1 Introduction

Online harmful language does not always look hostile (ElSherief et al., 2021). In Chinese short videos, CPCL often appears as sympathy, praise, advice, or moral concern, while implicitly positioning vulnerable groups as helpless, backward, dependent, or inferior (Glick and Fiske, 1996; Fiske et al., 2002; Sap et al., 2020). It may target people with disabilities, women, the elderly, children, single-parent families, and low-income populations (Wang et al., 2023, 2025). The harm lies less in a single offensive word than in the social relation created by the utterance (Sap et al., 2017, 2020).

This makes CPCL closer to pragmatic language understanding than to keyword-based toxicity detection (Davidson et al., 2017; ElSherief et al., 2021). A phrase that appears caring may become condescending when it repeatedly frames the target as helpless, dependent, backward, or grateful. Conversely, a video may quote discriminatory language in order to criticize it (Röttger et al., 2021). The label therefore depends on pragmatic stance, discourse role, speaker attitude, target positioning, and contextual framing rather than isolated offensive terms (Mohammad et al., 2016; Sap et al., 2017).

Short videos provide useful context, but they also blur the modeling target (Kiela et al., 2020; Gomez et al., 2020). If every evidence source is treated as an equal decision source, a model may learn shortcuts from scenes, tones, or facial expressions that correlate with vulnerable groups without capturing condescending stance. For CPCL, the decision should remain centered on language. ASR transcripts record what is said; comments reveal how audiences interpret or contest the video; and visual, acoustic, and facial cues help resolve whether an expression is sincere, ironic, mocking, pitying, or staged (Gomez et al., 2020). These nonverbal cues help interpret stance, but usually do not define CPCL independently.

We therefore start the model from textual evidence. ASR transcripts and comments are encoded first; visual, acoustic, and facial cues are projected to the same hidden size and combined with semantic knowledge evidence in a text-centric evidence classifier.

The paper makes four contributions. First, it formulates Chinese video-associated CPCL detection as contextual stance interpretation, where ASR transcripts provide primary linguistic evidence and comments provide post-publication social feedback

when available. Second, it extends PCLMM (Wang et al., 2025) to PCLMMPLUS with additional CPCL-oriented videos and 103,000 top-level comments, enabling controlled study of social feedback in Chinese video CPCL detection. Third, it evaluates a text-centric evidence classifier that organizes ASR, comments, nonverbal context, and Qwen-based semantic evidence around a shared textual decision state. Fourth, it reports ablations, robustness checks, group diagnostics, and qualitative cases showing that strong text baselines are competitive, nonverbal evidence is useful as contextual support, comments are informative but noise-sensitive, and Qwen-based semantic evidence improves sample-level representation without proving symbolic graph reasoning.

2 Related Work

PCL and CPCL detection. PCL has mainly been studied as text classification, including the "Don't Patronize Me!" dataset (Pérez-Almendros et al., 2020), condescension in conversational context (Wang and Potts, 2019), Chinese CCPC (Wang et al., 2023), and PclGPT (Wang et al., 2024b). These studies show that condescension is contextual and stance-dependent, so it cannot be reduced to sentiment classification or offensive keyword detection (Davidson et al., 2017; Röttger et al., 2021). PCLMM (Wang et al., 2025) extends CPCL detection to Chinese videos, but prior framing is closer to multimodal video classification. Our work keeps the video context while reformulating the task as text-centric contextual stance understanding.

Multimodal harmful content detection. Multimodal harmful-content datasets such as HateMM (Das et al., 2023), MultiHateClip (Wang et al., 2024a), MM-HSD (Céspedes-Sarrias et al., 2025), HVGuard (Jing et al., 2025), Hateful Memes (Kiela et al., 2020), and multimodal-publication hate speech data (Gomez et al., 2020) show that text, image, audio, and video can complement one another. Much of this work focuses on explicit hate speech, general video semantics, or parallel evidence sources. CPCL differs because harm is often hidden under politeness, sympathy, or concern. Treating modalities as equal voters can be risky: visual or acoustic patterns may correlate with social stereotypes without expressing condescending stance. We therefore model nonverbal streams as auxiliary contextual evidence for textual interpretation.

Dataset	Videos	PCL	Non-PCL	Comments
PCLMM (Wang et al., 2025)	715	196	519	–
PCLMMPLUS	831	313	518	103K

Table 1: Dataset statistics in the final working index. PCLMMPLUS extends PCLMM with additional CPCL-oriented videos and large-scale user comments.

Social feedback and knowledge-enhanced text classification. User comments can expose how audiences interpret or contest a video, but they are not labels and may contain sarcasm, bias, or off-topic discussion (Obadimu et al., 2019; Miyazaki et al., 2024; Bertaglia et al., 2024). Knowledge-enhanced methods help models reason beyond local text patterns; we build on sentiment knowledge graph modeling (Zhao and Yu, 2021) and use semantic knowledge evidence for implicit attitude reasoning. This is important because CPCL is not equivalent to negative sentiment: words such as care, help, poor, admirable, or resilient can be benign in one discourse and patronizing in another.

3 Dataset

3.1 PCLMM and PCLMMPLUS

PCLMM (Wang et al., 2025) is a Chinese video CPCL dataset from Bilibili with 715 videos, including 196 PCL and 519 non-PCL samples across six vulnerable groups. Labels are assigned at video level by three trained annotators and finalized by majority voting.

PCLMMPLUS keeps the PCLMM collection and extends it to 831 videos with additional CPCL-oriented samples and 103,000 top-level comments. The final index contains 313 PCL and 518 non-PCL samples, with 27 hours of video and about 3.927 million comment characters. We report the final working index used in all experiments rather than describing the resource as a simple arithmetic addition of newly collected videos, because duplicate removal, availability checks, and index-level filtering are applied before the split is fixed. The comments make it possible to study how audience responses relate to video-level CPCL labels, while still treating those responses as noisy evidence rather than annotation. Table 1 summarizes the dataset-level counts used in the paper.

3.2 Video-associated Textual Context

Each instance contains an ASR transcript, top-level comments, visual frame features, acoustic features, facial-expression features, and a CPCL

Split	Videos	PCL	Non-PCL
Train	600	227	373
Dev	66	25	41
Test	165	61	104

Table 2: PCLMMPLUS split by video ID. No video contributes evidence to more than one split.

label. Comments remain attached to video IDs and are not treated as separate labeled instances. Comments are post-publication social feedback: they support retrospective CPCL analysis after a video has received audience response, but they are not assumed to be available for immediate pre-publication moderation. This is still a meaningful content-governance setting, because viewers may recognize condescension even when the ASR alone is ambiguous. We therefore distinguish two use settings. A pre-publication or early-moderation setting should rely on ASR and intrinsic video evidence, optionally with knowledge evidence that can be built from the transcript. A post-publication retrospective setting can additionally use comments as noisy social feedback. We report ASR-only and no-comment settings as deployment-relevant checks. Comments are filtered to remove duplicates, emoji-only replies, and obvious spam, but they remain noisy evidence rather than labels. The split is by video ID, so comments and all feature streams from the same video remain in the same partition. Table 2 reports the working split used for the main experiments.

The test set keeps all six vulnerable-group categories: children, disabled people, the elderly, low-income people, single-parent families, and women. The number of positive test samples differs across groups, so we avoid drawing conclusions from accuracy alone and later report group-level diagnostics.

All experiments use the same video-level sample set, so performance differences across settings reflect the evidence sources supplied to the classifier rather than changes in the evaluated split.

4 Method

4.1 Problem Formulation

For each instance, let T be the ASR transcript, $C = \{c_1, \ldots, c_n\}$ comments, V visual evidence, A acoustic evidence, F facial evidence, and K a sentiment knowledge graph. The goal is

$$p(y = 1 \mid T, C, V, A, F, K), \tag{1}$$

where $y = 1$ indicates CPCL. Because CPCL is expressed mainly through language, we build an enhanced textual representation and use auxiliary evidence only to support interpretation of T and C.

The implementation separates evidence by role. ASR records what is said; comments record audience response; visual, audio, and face features describe delivery and framing; and knowledge evidence adds affective or group-related cues. The model first builds a textual state from transcript and comments, then restricts other sources to residual updates of that state.

4.2 Textual Context Encoder

ASR transcripts are produced by extracting video audio with FFmpeg (Developers, 2024), transcribing it with Whisper, and encoding the Chinese transcript with RoBERTa-wwm-ext (Cui et al., 2020; Wang et al., 2025). Reported runs use these pre-extracted ASR features projected into the hidden space rather than training a large PLM end to end:

$$H_T = \mathrm{Proj}(f_T), \tag{2}$$

where $H_T \in \mathbb{R}^{L \times d}$, $d = 768$, and L may be one for vector-level pre-extracted text features. The downstream model trains the projection layers and evidence classifier.

We use this projected ASR representation as the alignment target for the other evidence streams, whose lengths and reliability vary across videos.

4.3 Social Comment Encoder

User comments are encoded as social textual feedback. Individual comments are never treated as separately labeled training instances; they remain attached to their video ID. For efficiency and reproducibility, the reported implementation uses a lightweight hashing-based comment encoder with attention pooling over up to 16 non-empty top-level comments:

$$\alpha_i = \mathrm{softmax}(w^\top \tanh(W h_i)), \tag{3}$$

$$z_C = \sum_i \alpha_i h_i. \tag{4}$$

The 16-comment cap follows the fixed preprocessing used for all comment-aware models; videos with fewer valid comments use all available non-empty comments. The aggregated vector is used as textual context rather than as a separate prediction score. The branch is intentionally simple: the

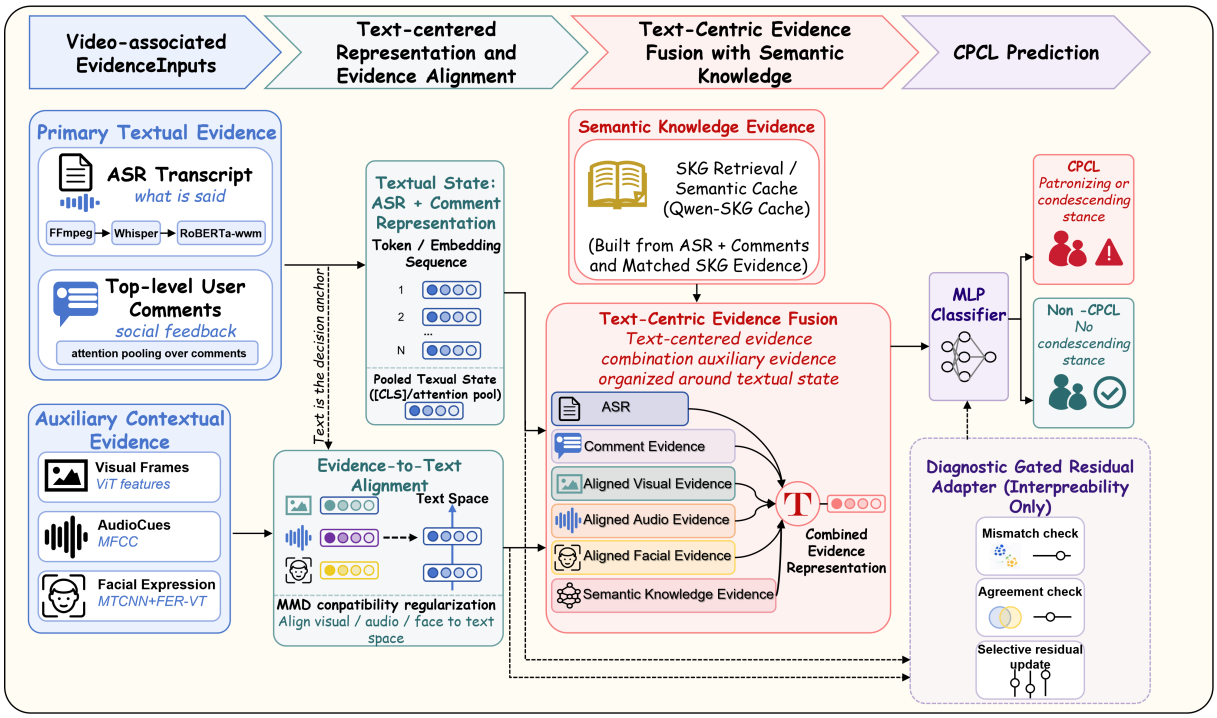

Figure 1: Text-centric evidence model for video-associated CPCL detection. Textual, nonverbal, and semantic evidence are projected into a shared hidden space before classification; the gated residual path is used as an interpretability diagnostic.

experiment tests whether video-level comment aggregation helps, not whether a stronger comment encoder would be optimal.

Some comments call out condescension, some repeat the harmful stance, and others discuss background events or other viewers. Attention pooling gives the model a limited way to downweight irrelevant comments, but audience response is still treated as input text, not supervision.

4.4 Auxiliary Evidence Encoding

For nonverbal evidence, video frames use ViT features (Dosovitskiy et al., 2020), faces use MTCNN (Zhang et al., 2016) and FER-VT (Huang et al., 2021), and audio uses FFmpeg (Developers, 2024) with acoustic features such as MFCC. Each stream is projected to the textual hidden size:

$$E_v = W_v \mathrm{VisualEncoder}(V), \quad (5)$$

$$E_a = W_a \mathrm{AcousticEncoder}(A), \quad (6)$$

$$E_f = W_f \mathrm{FacialEncoder}(F). \quad (7)$$

All visual, acoustic, and facial streams are projected to $d = 768$ before evidence fusion. These streams are not assumed to define CPCL alone; they provide context for reading textual stance.

A sympathetic sentence can become patronizing through tone, facial affect, or the way a target group is visually staged. A frame or facial expression without the utterance is much less informative. For this reason, these streams enter through the evidence adapter rather than through a separate video classifier.

4.5 Evidence-to-Text Alignment

Before evidence is combined with the textual representation, each auxiliary stream is aligned to the transcript representation:

$$E_m^{align} = \mathrm{Align}(E_m, H_T), \quad m \in \{v, a, f\}. \quad (8)$$

Text is the alignment anchor because CPCL is expressed through language. We regularize evidence-text compatibility with Maximum Mean Discrepancy (MMD):

$$\mathcal{L}_{\mathrm{MMD}} = \sum_{m \in \{v,a,f\}} \mathrm{MMD}(E_m^{align}, H_T). \quad (9)$$

MMD keeps auxiliary evidence close enough to the textual space for residual evidence updates without making modalities interchangeable.

This loss is a compatibility regularizer. It keeps projected evidence close enough to the textual state for residual updates to be numerically stable, without claiming that nonverbal streams have the same semantics as text.

4.6 Text-Centric Evidence Classifier

After alignment, the model combines the pooled textual state, comment vector, aligned nonverbal evidence, and knowledge vector before the final MLP:

$$z = [h_T; z_C; e_v; e_a; e_f; z_K], \quad \hat{y} = \sigma(\mathrm{MLP}(z)). \tag{10}$$

We also inspect a gated residual adapter as diagnostic evidence about how different sources interact with the textual state:

$$U_m = [H_T; E_m^{align}; H_T - E_m^{align}; H_T \odot E_m^{align}], \tag{11}$$

$$g_m = \sigma(W_m U_m), \tag{12}$$

$$\Delta H_m = g_m \odot E_m^{align}. \tag{13}$$

The enhanced textual representation is then

$$H'_T = \mathrm{LayerNorm}(H_T + \sum_m \Delta H_m + \Delta H_C + \Delta H_K), \tag{14}$$

where ΔH_C and ΔH_K are comment and knowledge updates.

The difference term $H_T - E_m^{align}$ gives the gate access to mismatch, the product term gives it access to agreement, and the sigmoid controls the size of the residual update. We use gate values only as an interpretability diagnostic, not as the main source of the final score.

4.7 Semantic Evidence for Implicit Attitude Cues

We treat the Qwen-SKG branch as semantic evidence rather than as proof of symbolic graph reasoning. Knowledge evidence is used because positive words can still be patronizing in a pitying frame. The lightweight SKG contains attribute–sentiment–polarity triples from Chinese PCL-related corpora (Wang et al., 2023; Zhao and Yu, 2021). The Qwen-SKG setting uses an offline cache built before model training from ASR transcripts, top-level comments, and substring-matched SKG triples. We encode the augmented text with a local Qwen3-Embedding-0.6B model (Zhang et al., 2025), keep one fixed vector per video, and use no gold label or prediction information in cache construction. The resulting z_K is projected into the hidden space and combined with the textual and nonverbal evidence vectors.

The lightweight graph can supply affective cues, but it is sparse and brittle for indirect stance. The cache is less interpretable but gives a richer sample-level vector. We keep both settings in the experiments because they answer different questions: whether sparse symbolic evidence is enough, and whether richer semantic evidence is useful.

The knowledge branch is included because CPCL is not equivalent to negative sentiment. Words such as care, help, or poor can be benign in one discourse and condescending in another. We therefore compare lightweight SKG, shuffled/random SKG, Qwen embedding-only evidence, and the full Qwen-SKG cache in the experiments.

4.8 Training Objective

The classifier predicts from the selected text-centered evidence representation z:

$$\hat{y} = \sigma(\mathrm{MLP}(z)). \tag{15}$$

The total training objective combines classification loss and alignment regularization:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{\mathrm{mmd}} \mathcal{L}_{\mathrm{MMD}}. \tag{16}$$

We use focal loss for $\mathcal{L}_{cls}$ when training on imbalanced data, following the positive-class recall needs of CPCL moderation.

5 Experiments

5.1 Experimental Setup

We evaluate on PCLMM for comparability and on PCLMMPLUS for the proposed text-centric setting. Main PCLMMPLUS results use five seeds: 13, 21, 42, 87, and 100. Thresholds are tuned only on development data and then fixed for test evaluation. Training uses AdamW, focal loss ($\alpha = 0.75, \gamma = 2.0$), batch size 8, learning rate 10^{-4}, weight decay 10^{-3}, gradient clipping at 1.0, and early checkpoint selection on development macro-F1; MMD runs use $\lambda_{\mathrm{mmd}} = 0.05$ with an RBF-kernel compatibility loss. ASR, visual, acoustic, and facial features are pre-extracted; downstream training updates projection and classifier layers. All tables are computed from video-level predictions, not comment-level rows.

We report accuracy, macro-F1, positive-class recall, precision, AUC, and PR-AUC. Macro-F1 is the primary thresholded metric, while recall and ranking metrics matter because moderation systems may tune thresholds according to review capacity. The Qwen-SKG cache is fixed before training, contains one evidence file for each of the 831 videos, and is constructed without labels.

Model	Acc.	$F1_m$	Recall	Prec.
BERT-PCL (Wang et al., 2023)	0.7972	0.7113	0.5294	0.5806
LLM TXT check*	0.7369	0.6653	0.3654	0.8462
VideoMAE (Tong et al., 2022)	0.7778	0.7090	0.5250	0.6176
MultiPCL (Wang et al., 2025)	0.8309	0.7978	0.7632	0.6744
CPCLDetector	0.8382	0.8130	0.8878	0.7317

Table 3: Representative baselines and external checks. * LLM TXT is an external transcript-only hard-label check on PCLMMPLUS, not a trained or directly comparable video model.

Baseline scope. The same-split baselines test whether the final model merely benefits from weak transcript encoders or generic video features. MacBERT and RoBERTa-wwm-ext provide strong text references, while VideoMAE+MacBERT tests a generic video-text baseline. These do not exhaust modern video-LLM systems, but they directly target the main comparison needed for this paper.

5.2 RQ1: Is CPCL Detection Text-Centered?

Table 3 reports representative baselines and external checks. The LLM TXT row is an external transcript-only hard-label check on PCLMMPLUS, while the remaining baselines are reported for comparison with prior PCLMM work. Its high precision and low recall show that generic prompt-based hard labeling is cautious but misses many CPCL cases, so we use it as a sanity check rather than as a directly comparable trained video model.

5.3 RQ2–RQ4: Contextual Evidence, Comments, and Knowledge

Table 4 reports same-split PCLMMPLUS results. The fine-tuned RoBERTa-wwm ASR+comments baseline is already strong, reaching 0.8300 macro-F1. The final text-centric evidence model adds aligned nonverbal evidence and Qwen-SKG semantic evidence, reaching 0.8497 macro-F1 and 0.9296 AUC. We treat this as a modest same-split gain rather than a dominance claim because we do not report paired prediction-level significance tests. The lower VideoMAE+MacBERT result suggests that generic video-text fusion is not sufficient for CPCL without text-grounded evidence organization.

Table 5 separates several sources that could explain the Qwen-SKG gain under a component-ablation protocol. Qwen semantic encoding improves over ASR and ASR+comments alone, and comments remain useful in the Qwen setting. Matched SKG triples help most clearly when comments are absent, while the full Qwen-SKG cache is close to the Qwen-ASR+comments and random-SKG settings. We therefore interpret Qwen-SKG as sample-level semantic evidence, not as proof that symbolic graph structure alone drives the improvement.

Model	Acc.	$F1_m$	Rec.	Prec.	AUC	PR
MacBERT ASR	0.7915 (.0259)	0.7823 (.0214)	0.8066 (.0901)	0.6958 (.0692)	0.8746 (.0084)	0.8110 (.0180)
MacBERT ASR+Com.	0.8424 (.0113)	0.8269 (.0186)	0.7475 (.0924)	0.8179 (.0463)	0.9102 (.0102)	0.8691 (.0253)
RoBERTa ASR	0.8061 (.0364)	0.7987 (.0326)	0.8361 (.0518)	0.7074 (.0676)	0.8861 (.0070)	0.8335 (.0132)
RoBERTa ASR+Com.	0.8400 (.0190)	0.8300 (.0166)	0.8098 (.0500)	0.7757 (.0610)	0.9057 (.0082)	0.8614 (.0128)
VideoMAE+MacBERT MM	0.7006 (.0308)	0.6769 (.0159)	0.5967 (.0932)	0.6108 (.0769)	0.7669 (.0200)	0.6588 (.0233)
Full model	0.8109 (.0387)	0.7968 (.0354)	0.7443 (.0739)	0.7594 (.0932)	0.9048 (.0091)	0.8692 (.0110)
Full+Qwen-SKG	**0.8618** (.0027)	**0.8497** (.0053)	0.7869 (.0602)	**0.8353** (.0475)	**0.9296** (.0087)	**0.8922** (.0115)

Table 4: Main same-split supervised results on PCLMMPLUS. Parentheses show five-seed standard deviation. RoBERTa denotes RoBERTa-wwm-ext; VideoMAE+MacBERT MM is the multimodal baseline; Full+Qwen-SKG is the final text-centric evidence model.

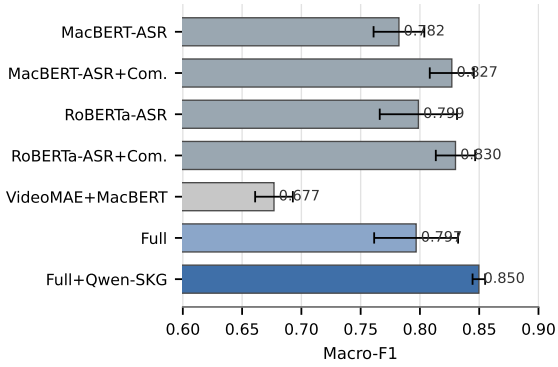

Figure 2: Main same-split Macro-F1 comparison on PCLMMPLUS. This figure visualizes the Macro-F1 column of Table 4; error bars denote five-seed standard deviation.

Protocol note. Tables 4, 5, and 7 answer different diagnostic questions: main same-split supervised performance, Qwen/SKG component diagnostics, and fixed-checkpoint robustness perturbations. The all-evidence row in Table 7 is therefore a protocol reference, not a replacement for the Table 4 final score.

5.4 Additional Same-Split Diagnostics

The same-split baselines in Table 4 make the comparison more demanding than the earlier PCLMM-only baselines. Fine-tuned text models are strong, so the final result should not be read as an easy gain over weak text encoders. The added value comes from evidence that those baselines do not use: aligned nonverbal context and cached semantic evidence.

The VideoMAE+MacBERT row provides a

Setting	$F1_m$	Recall	AUC	PR
ASR-only	0.7739 (.0131)	0.6689 (.0781)	0.8802 (.0103)	0.8354 (.0149)
ASR+Com.	0.7821 (.0180)	0.7213 (.1050)	0.8863 (.0100)	0.8456 (.0104)
Qwen-ASR	0.8026 (.0257)	0.6918 (.0639)	0.9113 (.0197)	0.8649 (.0374)
Qwen-ASR+Com.	**0.8345** (.0251)	0.7607 (.0720)	0.9366 (.0101)	0.9008 (.0133)
Qwen-SKG-ASR	0.8281 (.0169)	**0.7672** (.0511)	0.9252 (.0104)	0.8904 (.0146)
Full Qwen-SKG	0.8331 (.0174)	0.7279 (.0575)	**0.9378** (.0102)	**0.9017** (.0140)
Random-SKG	0.8313 (.0138)	0.7639 (.0620)	0.9335 (.0089)	0.8959 (.0103)

Table 5: Qwen-SKG component diagnostics, not a replacement for the final model score in Table 4. Parentheses show five-seed standard deviation. Random-SKG uses randomly matched SKG triples; these results support a conservative interpretation of Qwen-SKG as semantic evidence rather than isolated graph reasoning.

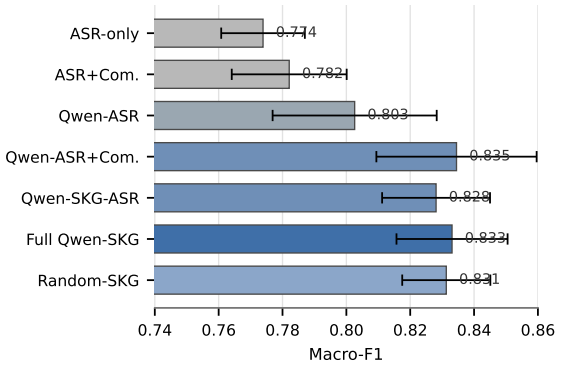

Figure 3: Macro-F1 diagnostics for Qwen-based semantic evidence. This figure visualizes the Macro-F1 column of Table 5; error bars denote five-seed standard deviation.

same-split multimodal reference. Its lower macro-F1 does not mean video features are unhelpful in general; it indicates that generic video representation is weaker than nonverbal evidence organized around the textual decision.

The Qwen component table clarifies the knowledge branch. Qwen-ASR improves over ASR-only, Qwen-ASR+comments improves further, and matched triples help most clearly when comments are absent. Because Full Qwen-SKG remains close to random-SKG, we report Qwen-SKG as semantic evidence construction rather than as a standalone graph reasoning module.

Finally, ASR-only and ASR+comments rows mark two practical boundaries: an intrinsic text setting before social feedback is available, and a post-publication setting where comments provide additional but noisy interpretation evidence.

5.5 Qualitative and Error Analysis

Table 6 presents four translated and lightly condensed test-set cases selected from a Qwen-SKG diagnostic model. They illustrate clear ASR evidence, non-CPCL self-narration, stance reversal, and ASR/comment mismatch; these patterns are not intended as an exhaustive error distribution.

5.6 Robustness and Group Diagnostics

Table 7 uses ✓ and × to show which evidence sources are enabled. Sparse, random, or noisy comments leave this checkpoint largely unchanged, suggesting that the lightweight comment branch is cautious; stronger ASR+comments and Qwen comment settings still show that comments can help when encoded by stronger textual or semantic representations. Disabling all nonverbal evidence causes larger drops.

Table 8 reports group-wise diagnostics for the final Full+Qwen-SKG model and marks groups with fewer than 10 positive test examples. These values are exploratory rather than stable group-level difficulty rankings.

6 Discussion

The experiments do not show that text alone is enough. They show that the transcript is where the hardest decision is made. Strong ASR and ASR+comments baselines are competitive, and the best result comes when textual evidence is combined with nonverbal context and semantic evidence. This fits the annotation problem: CPCL depends on how a group is described, positioned, or quoted, while tone, face, scene, and audience reaction help decide how that description should be read.

Comments behave unevenly for the same reason. They are useful because viewers have already interpreted the video, but they may also contain anger, stereotypes, or drifted discussion. We therefore treat comments as noisy post-publication evidence, not supervision. The fusion results also favor a practical architecture: a compact classifier over compatible text-centered evidence vectors is enough, while gated variants are mainly diagnostic. The Qwen results similarly point to sample-level semantic enrichment; matched triples can help, but their symbolic structure is not isolated as the main driver.

The additional diagnostics help delimit these conclusions. The same-split text baselines show

Case	Gold / Pred.	ASR evidence	Social feedback	Interpretation
TP: `low_income144`	CPCL / CPCL, 0.948	The speaker repeatedly attacks a worker's clothing and cleanliness, framing him as dirty and socially inappropriate.	Viewers notice the conflict and the shift of blame away from living or working conditions.	Explicit derogatory framing in the transcript supports the positive prediction; comments mainly provide confirming social feedback.
TN: `disabled104`	non-CPCL / non-CPCL, 0.105	A disabled creator narrates her own work, uncertainty, and self-development after becoming a full-time creator.	Comments are mostly supportive or share similar life experiences.	Vulnerable-group mention alone is not CPCL; the discourse role is self-narration rather than condescending positioning.
FP: `low_income50`	non-CPCL / CPCL, 0.846	The speaker reports age and gender discrimination in hiring and criticizes company profit logic.	Viewers discuss labor anxiety, resistance, and employment pressure.	The model likely confuses reported discriminatory content with endorsement, illustrating stance-reversal risk.
FN: `women45`	CPCL / non-CPCL, 0.485	The ASR transcript is mostly repeated promotional language, so the relevant harmful scene is poorly represented in text.	Comments discuss harassment-related context, but only indirectly reconstruct the video situation.	When ASR under-specifies the stance and comments are indirect, the text-centered model can miss CPCL.

Table 6: Qualitative case study of text-centric CPCL detection. ASR and comment evidence is translated and lightly condensed from the original Chinese materials. Probabilities are mean test probabilities over five seeds for a Qwen-SKG diagnostic model. The cases illustrate representative evidence patterns rather than exhaustive error categories.

Setting	Cmt.	Vis.	Aud.	Face	Mode	Acc.	Macro-F1	Recall	AUC	ΔF1
All evidence	✓	✓	✓	✓	Full	0.8000	0.7861	0.7377	0.8814	–
No comments	×	✓	✓	✓	None	0.7879	0.7732	0.7213	0.8712	-0.0129
One comment	✓	✓	✓	✓	One	0.8000	0.7861	0.7377	0.8813	0.0000
Random comments	✓	✓	✓	✓	Random	0.8000	0.7861	0.7377	0.8805	0.0000
Noisy comments	✓	✓	✓	✓	Noisy	0.8000	0.7861	0.7377	0.8814	0.0000
No visual	✓	×	✓	✓	Full	0.7818	0.7523	0.5902	0.8628	-0.0338
No audio	✓	✓	×	✓	Full	0.7758	0.7658	0.7705	0.8751	-0.0203
No face	✓	✓	✓	×	Full	0.7394	0.7213	0.6557	0.8275	-0.0648
No non-text	✓	×	×	×	Full	0.6848	0.6786	0.7377	0.8002	-0.1075

Table 7: Robustness diagnostics. ✓ and × indicate enabled evidence sources; ΔF1 is relative to all evidence.

Group	#Test	#CPCL	Low #CPCL	Macro-F1	Recall	Prec.	AUC
children	32	13	×	0.8032	0.6769	0.8605	0.9061
disabled	34	8	✓	0.9063	0.8000	0.9270	0.9760
elderly	20	7	✓	0.8128	0.8857	0.6981	0.9560
low-income	28	11	×	0.7919	0.8000	0.7223	0.8770
single-parent	22	9	✓	0.8356	0.7333	0.8798	0.9282
women	29	13	×	0.9221	0.8615	0.9679	0.9558

Table 8: Group-wise diagnostics for the final Full+Qwen-SKG model. #Test and #CPCL denote test videos and positive CPCL videos. ✓ in Low #CPCL indicates fewer than 10 positive test examples; results are diagnostic rather than stable group rankings.

that the gains are not obtained by comparing against weak transcript encoders. The VideoMAE+MacBERT reference, in contrast, shows that simply adding a generic video representation to a strong text encoder is not sufficient for this task. This contrast separates two claims that are easy to conflate: video context can matter for CPCL, but it matters most when it conditions how textual stance is read.

The robustness and qualitative cases point to the same boundary. Removing all non-text evidence hurts more than perturbing comments in the robustness checkpoint, while ASR+comments and Qwen comment settings show that comments can still add value when represented with stronger semantic encoders. Quotation, self-narration, and critical discussion can contain CPCL-like surface words without endorsing a condescending stance; conversely, a positive video may hide condescension in repeated framing of dependence or gratitude. Taken together, the current evidence supports a specific claim: under PCLMMPLUS, CPCL prediction is most reliable when linguistic stance remains the decision center and comments, nonverbal cues, and semantic caches are treated as contextual evidence.

7 Conclusion

We studied Chinese video-associated CPCL detection as contextual stance interpretation grounded in text. PCLMMPLUS extends PCLMM to 831 videos and adds 103,000 comments for studying social feedback. Under this setting, strong textual baselines are competitive, nonverbal context helps when organized around the textual decision, and Qwen-based semantic evidence improves sample-level representation. The final text-centric evidence model reaches 0.8497 macro-F1 and 0.9296 AUC.

The experiments also delimit the claim: comments are post-publication feedback, Qwen-SKG should be interpreted as semantic evidence rather than isolated graph reasoning, and stronger video-LLM baselines remain future diagnostics. More broadly, the findings suggest that polite-but-harmful language in short videos should be modeled through linguistic stance, with social and multimodal context used cautiously as supporting evidence. This framing also gives a concrete evaluation template: compare strong text-only settings, comment-aware settings, and contextual-evidence variants under the same split before attributing gains to multimodal fusion.

Limitations

This study includes group-wise diagnostics for one Qwen-SKG variant, but not a full per-group comparison for every baseline and ablation. The dataset scale remains limited, especially for positive CPCL samples within some vulnerable groups, and video-level labels do not localize which utterance, frame, or comment expresses condescension. A fuller dataset release should include the incremental sampling log, annotation guideline, inter-annotator agreement, and label reconciliation details for the added videos. ASR errors can weaken the textual backbone, while nonverbal streams may introduce shortcuts if not checked against text-centered baselines. Comments are socially informative but noisy and platform-specific; they may encode bias, sarcasm, or thread-level discussion rather than video-grounded interpretation. The test set is small, and five-seed variance does not replace paired prediction-level significance testing; future work should report bootstrap confidence intervals or approximate-randomization tests against the strongest text baselines. The LLM TXT check is based on binary outputs rather than calibrated probabilities and lacks full public prompt/model reproducibility, so we treat it only as an external transcript-only sanity check. Stronger video-LLM baselines and broader per-group evaluation remain important future diagnostics.

Ethical Considerations

CPCL detection concerns vulnerable groups, including children, disabled people, elderly people, women, low-income people, and single-parent families. Both false positives and false negatives can cause harm. False positives may suppress self-narration, advocacy, criticism, or supportive discussion by members of vulnerable groups, while false negatives may leave subtle condescension unflagged. The model should therefore support human-in-the-loop moderation and research analysis rather than replace human judgment or produce punitive decisions. Group-wise diagnostics are exploratory and should not be used to profile communities or rank group difficulty.

The comment branch requires particular care. Comments are public post-publication social feedback, but commenters may not expect their text to be used for research, and comments can include stereotypes, harassment, private experiences, or other sensitive disclosures. We treat comments as noisy evidence rather than labels or crowd judgments. If released, user identifiers, profile links, avatars, and other profile-level information should be removed. Raw comments or video identifiers should be distributed only when permitted by platform policy and after de-identification; otherwise, release should prioritize processed features, labels, splits, and anonymized text fields.

Because the data may involve minors or other vulnerable individuals, the dataset should not be used for autonomous surveillance, profiling, or fully automated moderation. Any deployment should include human review, contextual inspection, and an appeal or correction mechanism. Future releases should document data access conditions, deletion or takedown handling when feasible, and annotator protection procedures for exposure to harmful or stigmatizing content.

References

Thales Bertaglia, Catalina Goanta, and Adriana Iamnitchi. 2024. The monetisation of toxicity: Analysing youtube content creators and controversy-driven engagement. In *4th International Workshop on OPEN CHALLENGES IN ONLINE SOCIAL NETWORKS*, HT '24, page 1–9. ACM.

Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. 2025. Mm-hsd: Multi-modal hate speech detection in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2546–2555.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515.

PéFFmpeg Developers. 2024. Ffmpeg. *https://ffmpeg.org/*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias

Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan T. Fiske, Amy J. C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6):878–902.

Peter Glick and Susan T. Fiske. 1996. The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70(3):491–512.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. 2021. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 580:35–54.

Yiheng Jing, Mingming Zhang, Yong Zhuang, Jiacheng Guo, Juan Wang, Xiaoyang Xu, Wenzhe Yi, Keyan Guo, and Hongxin Hu. 2025. Hvguard: Utilizing multimodal large language models for hateful video detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Kunihiro Miyazaki, Takayuki Uchiba, Haewoon Kwak, Jisun An, and Kazutoshi Sasahara. 2024. The impact of toxic trolling comments on anti-vaccine youtube videos. *Scientific Reports*, 14(1):5088.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying toxicity within youtube video comment. In *International conference on social computing, Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 214–223. Springer.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. *arXiv preprint arXiv:2011.08320*.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502.

Hongbo Wang, Mingda Li, Junyu Lu, Liang Yang, Hebin Xia, and Hongfei Lin. 2023. Ccpc: A hierarchical chinese corpus for patronizing and condescending language detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 640–652. Springer.

Hongbo Wang, LiMingDa LiMingDa, Junyu Lu, Hebin Xia, Liang Yang, Bo Xu, Ruizhu Liu, and Hongfei Lin. 2024b. Pclgpt: A large language model for patronizing and condescending language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6913–6928.

Hongbo Wang, Junyu Lu, Yan Han, Kai Ma, Liang Yang, and Hongfei Lin. 2025. Towards patronizing and condescending language in chinese videos: A multimodal dataset and detector. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

Appendix

A Implementation and Feature Details

Table 9 summarizes the feature sources used by the reported models. The downstream classifier operates on pre-extracted evidence vectors and trains the projection layers and classifier parameters. The ASR pipeline follows the PCLMM preprocessing style: audio is extracted with FFmpeg, transcribed with Whisper, and encoded with RoBERTa-wwm-ext. All nonverbal evidence is projected to the shared hidden space before it is combined with textual and semantic evidence.

This cache-based setup separates feature extraction from supervised CPCL training. It allows the same video split and labels to be used when comparing strong text encoders, video-derived evidence, comment-aware variants, and Qwen semantic caches. The comparison is therefore focused on how evidence is represented and combined, rather than on changes in dataset construction or preprocessing order.

For reproduction, the intended path is file-level rather than description-only: feature caches define the inputs, per-seed output folders define the trained checkpoints and predictions, result summaries define the aggregated statistics, and paper-table files define the manuscript-facing presentation. This organization also makes it possible to rerun table export without rebuilding feature caches or retraining every model.

Component	Details
ASR transcript	Source: video audio. Extractor: FFmpeg, Whisper, RoBERTa-wwm-ext. Representation: transcript feature, 768-d. Use: projected to textual state.
Visual context	Source: video frames. Extractor: ViT. Representation: frame/video feature. Use: projected to 768-d evidence.
Acoustic context	Source: audio track. Extractor: MFCC. Representation: acoustic feature. Use: projected to 768-d evidence.
Facial context	Source: detected faces. Extractor: MTCNN, FER-VT. Representation: facial expression feature. Use: projected to 768-d evidence.
Comments	Source: top-level comments. Extractor: hashing encoder with attention pooling. Representation: social feedback vector. Use: aggregated at video level.
Semantic evidence	Source: ASR/comments/SKG text. Extractor: Qwen3-Embedding-0.6B. Representation: per-video cache, 1024-d. Use: projected as semantic evidence.

Table 9: Feature extraction and model input summary. The table reports the components used by the submitted experiments; it does not introduce additional training data or labels.

The comment branch uses at most 16 top-level comments under the fixed preprocessing used for comment-aware models. The Qwen semantic cache is constructed once before training and is indexed by video ID. It is not built from gold labels or model predictions.

Table 10 summarizes the main local artifacts used to reproduce the reported analyses. The paths are reported to make clear which files correspond to raw experiment outputs, summary tables, and fixed cache indexes. The table is descriptive only; it does not add extra data beyond the experiments reported in the paper.

This separation is useful for auditing the experimental chain. Per-seed outputs are the source of the reported means and standard deviations, result summaries aggregate those runs into reviewer-facing tables, and the Qwen cache indexes identify which fixed semantic vector is attached to each video. Keeping these artifacts separate makes it easier to check that feature construction, supervised training, and table export are not conflated.

The artifact map also distinguishes reproducibility files from data-release decisions. Local paths identify what was used to produce the reported tables, while public release should still follow the privacy and platform-policy constraints discussed

in Appendix H.

Artifact group	Example path	Use in the paper
Per-seed runs	`outputs/`	checkpoints, test metrics, and prediction files for individual seeds
Result summaries	`results/`	CSV summaries, audit reports, bootstrap diagnostics, and component analyses
Paper tables	`paper_tables/`	Markdown/LaTeX-ready tables used to prepare the manuscript
Qwen text caches	`features/QWEN_*`	fixed semantic evidence features for ASR, comments, and knowledge variants
Cache indexes	`data/index_qwen_*`	video-level indexes that attach cache features to the fixed split

Table 10: Artifact map for reproducibility. Paths are relative to the local experiment workspace and describe existing files rather than additional experiments.

B Training Protocols and Table Scope

Table 11 lists the common training and evaluation settings used by the text-centric experiments. All reported neural results use video-level predictions; comments are attached to the video and are not treated as independent labeled instances.

Setting	Value
Seeds	13, 21, 42, 87, 100
Optimizer	AdamW
Batch size	8
Learning rate	1×10^{-4}
Weight decay	1×10^{-3}
Gradient clipping	1.0
Focal loss	$\alpha = 0.75$, $\gamma = 2.0$
MMD weight	$\lambda_{\mathrm{mmd}} = 0.05$ when used
Threshold	selected on development split
Metrics	Acc., Macro-F1, Recall, Precision, AUC, PR-AUC

Table 11: Training and evaluation settings for the reported text-centric experiments.

The main hyperparameters were kept fixed across compared models whenever applicable. The learning rate, batch size, focal-loss settings, and MMD weight follow the submitted configuration files, while the classification threshold is selected on the development split only. Full hyperparameter sweeps are not reported in the current submission; future releases should include sensitivity curves for λ_{mmd}, decision threshold, comment count, and semantic evidence size.

Table 12 clarifies the scope of the main result tables. This distinction is useful because the main same-split results, Qwen component diagnostics, robustness perturbations, and group-wise diagnostics answer different questions.

Table	Protocol scope
Table 4	Main same-split supervised results on PCLMMPLUS; five seeds; primary performance comparison.
Table 5	Qwen/SKG component diagnostics; five seeds; attribution diagnostic, not the final model score.
Table 7	Evidence perturbation robustness with a fixed checkpoint/protocol; within-protocol reference only.
Table 8	Vulnerable-group diagnostics for the final Full+Qwen-SKG model; exploratory group-level analysis.

Table 12: Protocol map for the main tables. The diagnostics should not be read as repeated evaluations of the same checkpoint.

C Result Visualizations

The result visualizations do not introduce new experimental results. Figures 2 and 3 in the main paper visualize the Macro-F1 columns of Tables 4 and 5 to make the trends easier to inspect, while the tables retain the complete metric set.

D Qwen-SKG Cache Construction

The Qwen-SKG branch is implemented as a semantic evidence cache. For each video, the cache builder collects the configured text fields, matches up to 16 SKG triples by the configured matching rule, concatenates the selected evidence text, and encodes it with Qwen3-Embedding-0.6B. The full cache uses ASR, comments, and matched SKG triples; the random-SKG diagnostic replaces the matched triples with random triples using seed 2026. Metadata files report 831 indexed videos, label-free construction, top-$k = 16$ triples, and an average of 15.69 matched triples in the full rebuilt cache. The random-SKG cache uses 16 random triples per video.

Cache variant	ASR	Comments	SKG	Triple mode
Qwen-ASR	✓	×	×	–
Qwen-ASR+Com.	✓	✓	×	–
Qwen-SKG-ASR	✓	×	✓	matched
Full Qwen-SKG	✓	✓	✓	matched
Random-SKG	✓	✓	✓	random

Table 13: Qwen semantic-cache variants used in the component diagnostics. All variants are built before supervised training and do not use gold labels.

These diagnostics motivate the paper's conservative interpretation. Qwen encoding contributes much of the gain, and matched triples are most clearly useful when comments are absent. The small gap between Full Qwen-SKG and Random-

SKG does not support a claim that symbolic graph structure alone drives the result; it supports treating the cache as sample-level semantic enrichment.

The cache construction also separates semantic feature generation from supervised training. The ASR-only cache, ASR+comments cache, and ASR+comments-as-knowledge cache are indexed by video ID before model training begins. During supervised training, the classifier reads the corresponding cache vector and learns only how to combine it with the other evidence vectors. This separation is important for leakage control: the cache contains textual evidence and matched or randomized knowledge text, but not the gold CPCL label or any prediction from the downstream classifier.

E Robustness and Stability Diagnostics

The robustness visualization summarizes Table 7 and should be interpreted within that diagnostic protocol only; it is not a replacement for the main five-seed result in Table 4. Figure 4 shows the change in Macro-F1 relative to the all-evidence robustness setting. The unchanged random/noisy comment rows indicate limited sensitivity of this checkpoint to those perturbations, not that comments are universally reliable.

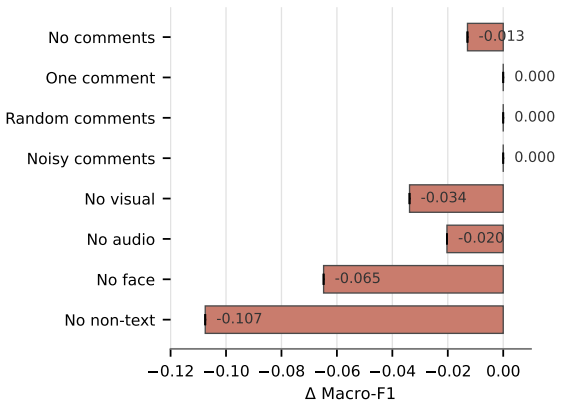

Figure 4: Robustness Δ Macro-F1 visual summary based on Table 7. Values are computed relative to the all-evidence setting within the robustness diagnostic protocol, not relative to the main final score in Table 4.

The main tables report five-seed means and standard deviations. In addition, existing prediction files allow paired bootstrap checks for the final Full+Qwen-SKG classifier against strong ASR+comments baselines. Table 14 summarizes 1000 paired bootstrap resamples over test videos, averaged across the five seeds. The intervals for Macro-F1 include zero, so we use the final improvement as numerical evidence under the current split rather than as a statistical dominance claim.

Comparison	Metric	Mean Δ	Mean 95% CI
Full+Qwen-SKG − RoBERTa ASR+Com.	Macro-F1	0.0114	[-0.0560, 0.0781]
Full+Qwen-SKG − RoBERTa ASR+Com.	AUC	0.0297	[-0.0151, 0.0773]
Full+Qwen-SKG − RoBERTa ASR+Com.	PR-AUC	0.0404	[-0.0293, 0.1140]
Full+Qwen-SKG − MacBERT ASR+Com.	Macro-F1	0.0145	[-0.0566, 0.0869]
Full+Qwen-SKG − MacBERT ASR+Com.	AUC	0.0252	[-0.0213, 0.0736]
Full+Qwen-SKG − MacBERT ASR+Com.	PR-AUC	0.0328	[-0.0383, 0.1095]

Table 14: Prediction-level paired bootstrap diagnostics from existing test predictions. Values are averaged across five seeds and should be read as stability diagnostics rather than definitive significance claims.

F Additional Comment-Availability Diagnostic

Table 15 reports an additional comment-availability diagnostic from the existing reviewer-response experiment set. This diagnostic uses a separate lightweight text setting, so the values should not be compared directly with the final model in Table 4. Its purpose is narrower: to check whether simple comment perturbations change the model behavior when comments are supplied through the lightweight comment path. The nearly identical rows under top-k, random, and keyword-removed comments indicate that this branch is conservative and does not strongly exploit fine-grained comment content in this diagnostic.

Setting	Acc.	Macro-F1	AUC	PR-AUC
ASR-only	0.6182	0.5979	0.7049	0.5853
All comments	0.6509	0.6131	0.6837	0.5795
Top-1 comment	0.6509	0.6131	0.6837	0.5795
Top-3 comments	0.6509	0.6131	0.6836	0.5795
Random comment	0.6509	0.6131	0.6837	0.5795
Keyword-removed	0.6509	0.6131	0.6837	0.5795

Table 15: Additional comment-availability diagnostic. All rows use ASR; comment rows vary the comment input. Values are five-seed means from the existing diagnostic run.

G Dataset and Qualitative Diagnostics

PCLMMPLUS is reported using the final working index used by all experiments. The index contains 831 videos and is not described as a simple arithmetic addition of newly collected videos because duplicate removal, availability checks, and index-level filtering are applied before the split is fixed. Detailed incremental sampling logs, annotation guidelines, agreement statistics, and label reconciliation notes should accompany a full dataset release when available.

The qualitative cases in the main paper are selected from existing prediction files and are shown

with English translations of Chinese ASR and comments. They are intended to illustrate recurring error patterns rather than estimate their frequency. The observed patterns include reported speech mistaken for endorsement, self-narration versus other-framing, praise or concern used as condescension, and ASR/comment mismatch. Group-wise scores in Table 8 are likewise exploratory: several groups contain fewer than 10 positive test videos, so the table should not be interpreted as a stable ranking of group difficulty.

H Data Release and Ethical Use

We recommend releasing processed features, split files, labels, and anonymized text fields as the default research artifact. Raw videos, direct video identifiers, or raw comments should be distributed only when permitted by platform policy and after de-identification. This release design supports reproducibility while reducing re-identification and privacy risks.

Material	Recommended release and mitigation
Video IDs/URLs	Release only if permitted by platform policy; reduce re-identification risk through access control or de-identification.
Raw comments	Avoid public release unless permitted; remove user identifiers and profile links, or provide anonymized text or processed features.
ASR transcripts	Release anonymized text when feasible; remove names or sensitive identifiers when detected.
Pre-extracted features	Preferred release artifact; supports reproducibility with lower user-identification risk.
Labels and splits	Release with video-level indexes that do not expose user profiles; preserve split integrity.
Qwen-SKG cache	Release if label-free; include metadata showing no gold labels or predictions were used.
Annotation guidelines	Release for transparency while removing annotator identifiers.

Table 16: Recommended data-release format and risk controls. The table is a release protocol, not a claim that all raw materials can be publicly distributed.

I External Qwen-VL API Reference

Table 17 reports an external zero-shot Qwen-VL API reference from the existing reviewer-response experiment set. The model is not trained on PCLMMPLUS and is not directly comparable to the supervised same-split models in Table 4. The run uses `qwen-vl-plus`, eight sampled frames per video, temperature 0, and three inputs: video, video+ASR, and video+ASR+comments. Metrics are computed over 153 successfully processed test videos; unprocessed samples are recorded in the API/data failure logs. Returned confidence scores are used for AUC and PR-AUC.

Input	Result summary
Video	$N = 153$; Acc. 0.5294; Macro-F1 0.5284; Recall 0.6491; Precision 0.4157; AUC 0.4743; PR-AUC 0.3567.
Video+ASR	$N = 153$; Acc. 0.4771; Macro-F1 0.4706; Recall 0.7895; Precision 0.3982; AUC 0.6118; PR-AUC 0.4353.
Video+ASR +Com.	$N = 153$; Acc. 0.5294; Macro-F1 0.5278; Recall 0.7895; Precision 0.4286; AUC 0.5716; PR-AUC 0.4114.

Table 17: External zero-shot Qwen-VL API reference on the PCLMMPLUS test split. The run has zero parse failures and 12 API/data failures per setting; the comment setting corresponds to post-publication retrospective analysis.

J Recommended Additional Diagnostics

Several additional diagnostics would strengthen future releases, but we do not report them as completed experiments here. The most useful next checks are broader video-LLM baselines beyond the single Qwen-VL reference above, paired bootstrap tests over all major ablations, MMD and threshold-sensitivity curves, comment-count or temporal comment analyses, stronger PLM-based comment encoders, ASR quality checks, and semantic top-k sweeps for Qwen-SKG attribution. These require additional controlled runs or manual checks, so no result numbers are assigned here.