PROOF OF CONCEPT OF ANALYZING MULTIMODAL PRAGMATIC MARKERS USING QWEN VL LLM

Hendi Pratama[1)], Stephani Priyanto Putri[2)]

1) Department of English, Faculty of Languages and Arts
Universitas Negeri Semarang, Semarang, Indonesia
hendipratama@mail.unnes.ac.id

Abstract

Pragmatic markers in spoken interaction are realized not only through lexical and prosodic resources but also through facial expression, gesture, and posture. The manual coding of such multimodal markers is labor-intensive and difficult to scale, which has constrained the systematic inclusion of multimodal evidence in pragmatic research. This study reports a proof of concept for a reproducible pipeline that supports the description and annotation of multimodal pragmatic markers using an open-weight Vision-Language Large Language Model (VL LLM), with politeness strategies (Brown & Levinson, 1987) deployed as a concrete instantiation of multimodal pragmatic markers. The pipeline was implemented on a Google Colab Pro environment with an NVIDIA L4 GPU and comprises four sequential modules. Audio is transcribed at the word level using Whisper Large-v3 with inter-word pauses of 0.5 seconds or greater marked inline. Video frames are extracted at three-second intervals using ffmpeg. Transcribed speech is segmented into propositions using hybrid punctuation-plus-LLM rules and then aggregated into speech acts via rolling-window LLM decisions. For each speech act, three sampled frames are submitted to Qwen3-VL-4B-Instruct with a structured prompt that elicits facial, gestural, and emotional descriptions and assigns Likert-scaled scores (0–5) on the four Brown and Levinson politeness superstrategies, accompanied by an evidence-grounded rationale. To accommodate the 22 GB usable VRAM ceiling, the speech recognition model is unloaded before the VL LLM is loaded. Application to a twelve-minute public-speaking sample yielded 160 propositions aggregated into 40 speech acts with full annotation coverage and no inference errors. Bald On-Record emerged as the dominant superstrategy in 62.5% of units, consistent with the assertive declarative register of motivational public speaking, while multi-label scoring revealed substantial strategy blending across half of the corpus. Diagnostic flagging identified 20.0% of units as warranting expert review. The contribution of the study is methodological: the pipeline demonstrates technical feasibility and structural tractability for multimodal pragmatic annotation, providing a reproducible foundation for subsequent inter-rater validation and fine-tuning work.

Keywords: multimodal pragmatics, pragmatic markers, politeness theory, Vision-Language Model, Qwen-VL, proof of concept

Introduction

Pragmatic meaning in face-to-face communication is realized through a coordinated interplay of verbal and non-verbal signals. Speakers signal stance, manage interpersonal alignment, and frame illocutionary force through facial expression, gesture, posture, and prosody (Kendon, 2004; Kress, 2010; McNeill, 1992). Pragmatic markers, in particular, function as contextualization cues that organize discourse, regulate speaker-hearer relations, and encode affective and epistemic positioning (Aijmer, 2013; Schiffrin, 1987). When pragmatic markers are understood as multimodal phenomena, they include not only discourse particles such as well, you know, and actually, but also the gestural emblems, facial displays, and bodily orientations that co-occur with speech and shape its interpretation.

Among the most theoretically developed accounts of how speakers manage interpersonal alignment is the politeness framework articulated by Brown and Levinson (1987). The framework identifies four superstrategies for managing face-threatening acts: Bald On-Record (direct, unmitigated), Positive Politeness (oriented to the hearer's desire for approval and inclusion), Negative Politeness (oriented to

the hearer's desire for autonomy and non-imposition), and Off-Record (indirect, implicational). Politeness strategies are themselves a class of pragmatic markers, realized through coordinated verbal, prosodic, and visual resources. The framework therefore offers a theoretically rich anchor for demonstrating the analysis of multimodal pragmatic markers in concrete data.

Despite the theoretical consensus on the multimodal nature of pragmatic communication, empirical research has struggled to operationalize this insight at scale. The principal obstacle is labor cost. Manual coding of multimodal pragmatic markers requires expert annotators to view recorded interaction, identify relevant verbal and visual signals, transcribe them with precise temporal alignment, and assign functional labels following theoretically motivated schemes. A single hour of recorded interaction can require many hours of annotation time, and inter-rater reliability further increases the labor required. Consequently, multimodal pragmatic corpora have remained small, idiosyncratic across research teams, and difficult to compare across studies.

Recent advances in vision-language large language models (VL LLMs) introduce a possible affordance for addressing this bottleneck. Models such as the Qwen-VL family (Bai et al., 2025) integrate visual and textual reasoning within a single inference pass, accept image inputs alongside natural-language prompts, and produce structured outputs in formal English. Their capacity for grounded multimodal description suggests potential utility for linguistic and pragmatic research (Lin, 2025; Ostyakova et al., 2025; Yu et al., 2024). The empirical question, however, is whether such models can produce annotations of facial conduct, gesture, and politeness strategy that are sufficiently consistent, time-aligned, and theoretically defensible to support downstream pragmatic analysis.

This study reports a proof of concept (PoC) addressing that question. The contribution of the present work is methodological rather than empirical. The aim is not to advance a substantive claim about the politeness deployment of any particular speaker, but to demonstrate the feasibility of building a working pipeline that integrates speech recognition, vision-language modeling, prosodic analysis, and structured politeness annotation into a single reproducible workflow. The pipeline is implemented on a constrained hardware configuration commonly accessible to applied linguistics researchers (Google Colab Pro with an L4 GPU, 22 GB usable VRAM), uses only open-weight models, and produces outputs in formats suitable for downstream coding and statistical analysis.

Two research questions guide the present work. First, can a pipeline integrating ffmpeg-based frame extraction, Whisper-based speech recognition, and Qwen VL LLM inference be configured to operate within the 22 GB VRAM ceiling typical of consumer GPU environments? Second, do the resulting outputs exhibit sufficient structural consistency and theoretical coherence to support downstream rubric-based coding by human analysts? The remainder of this paper describes the pipeline architecture, reports application to a public-speaking video sample, presents the structure and characteristics of the outputs, and discusses implications for subsequent validation work.

Methodology

This proof-of-concept study adopts a descriptive computational pipeline design. The pipeline is implemented in Python within a Google Colab Pro environment with NVIDIA L4 GPU acceleration (22 GB usable VRAM out of 24 GB physical capacity). All inference uses greedy decoding to ensure reproducibility. The pipeline comprises four sequential modules described below.

Pipeline Architecture

The first module performs media preprocessing. Audio is extracted from the source video file using ffmpeg, resampled to 16 kHz mono, and saved as a WAV file. Video frames are extracted at fixed three-second temporal intervals using ffmpeg with JPEG quality 15, yielding a frame inventory that downstream modules can sample without re-decoding the source video.

The second module performs automatic speech recognition. The Whisper Large-v3 model (Radford et al., 2022) is loaded onto GPU memory and configured for English transcription with word-level timestamps enabled. Audio is segmented into overlapping chunks of 300 seconds with 10-second overlaps, and word-level outputs from each chunk are merged with boundary deduplication. Inter-word pauses exceeding 0.5 seconds are marked inline within the transcript using the convention (X.X), preserving prosodic boundary information for downstream multimodal analysis. Prosodic features (F0 and intensity contours) are extracted using Parselmouth (Jadoul et al., 2018). Upon completion of the speech recognition stage, the Whisper model is unloaded from GPU memory before the vision-language model is loaded; this sequential loading is essential for operating within the 22 GB VRAM ceiling.

The third module performs unit segmentation. Transcribed speech is first segmented into propositional units using a hybrid approach combining punctuation-based initial segmentation, comma-boundary splitting for utterances exceeding 30 words, and large-language-model refinement for ambiguous cases. Propositions are then aggregated into speech-act-level units through a rolling-window LLM decision procedure. For each candidate proposition, the model evaluates whether it continues the current speech act or initiates a new one, considering both topical coherence and pragmatic function shift. When initiating a new speech act, the model assigns a brief functional label (e.g., narrate Captain Swenson story, state learned insight). The aggregation procedure follows Searle's (1969) conception of the speech act as the minimal unit performing a single illocutionary function.

The fourth module performs vision-language annotation. The Qwen3-VL-4B-Instruct model (Bai et al., 2025) is loaded with bfloat16 precision, occupying approximately 9 GB of VRAM. For each speech act, three video frames are sampled at the unit's start, midpoint, and end timestamps, and submitted to the model alongside the verbal transcript with pause markers, the auto-generated functional label, and duration and word-count metadata. The model receives a system prompt containing definitions of the four Brown and Levinson superstrategies, each described in a single paragraph emphasizing typical realizations and pragmatic function. The user prompt requests five outputs: a face description, a gesture description, an emotion inference, Likert-scaled scores (0–5) for each of the four superstrategies, the dominant strategy assignment, and a paragraph-length rationale citing both verbal phrases and visual cues. Greedy decoding is used throughout (do_sample = False), with maximum generation length of 600 tokens. Raw model outputs are parsed using regular-expression patterns and exported to an Excel workbook with embedded frame thumbnails and a CSV backup.

Likert Scoring and Multi-Label Design

The Likert scale operationalizes scoring independence: each superstrategy is scored on its own 0–5 dimension without forcing exclusivity. A speech act may therefore exhibit substantial presence of multiple strategies simultaneously, capturing the strategy-blending common in natural discourse. This design choice abandons the single-label classification typical of earlier politeness annotation studies, which forces a categorical decision and obscures the layered nature of natural politeness work.

Diagnostic Quality Assessment

Seven diagnostic flags are computed post-hoc to identify edge cases warranting expert review: visual hallucination (the model describes non-existent visual content), text artifact (transcript contains broadcast metadata such as translator credits), short act (duration under 3 seconds with fewer than 8 words), long act (duration exceeding 60 seconds, suggesting potential over-aggregation), uncertain dominant (tied maximum Likert scores across categories), all-low scores (maximum score ≤ 2), and minimal rationale (rationale text under 50 characters).

Sample Application

The pipeline was applied to a public-speaking video sample drawn from the publicly available TED archive. The sample comprises a single-speaker monologue of approximately twelve minutes in duration, recorded at 1280 × 720 resolution with stereo AAC audio at 44.1 kHz. This sample was chosen for three properties relevant to PoC validation: a single speaker without diarization complications, broadcast-quality audio-visual conditions, and rhetorical density typical of formal public-speaking discourse. The role of the sample is to demonstrate that the pipeline produces structured, theoretically coherent output of the intended form.

Findings and Discussion

Pipeline Execution and Operational Feasibility

The pipeline executed successfully on the twelve-minute sample within the constrained Google Colab Pro environment. Table 1 reports the technical specifications. The sequential loading of speech recognition and vision-language models, with the former unloaded prior to loading the latter, kept GPU memory utilization within the 22 GB ceiling throughout.

Table 1
Pipeline Specification

Component	Specification
Hardware platform	Google Colab Pro, NVIDIA L4 GPU (22 GB VRAM)
Speech recognition	OpenAI Whisper Large-v3, English, word-level timestamps
Vision-language model	Qwen3-VL-4B-Instruct (bf16, greedy decoding)
Frame extraction	ffmpeg, JPEG quality 15, three-second interval
Frame sampling per unit	Three frames (start, midpoint, end)
Prosodic analysis	Parselmouth (Praat) F0 and intensity, pitch floor 75 Hz
Pause marking convention	Inline (X.X) for gaps of 0.5 s or longer
Proposition segmentation	Punctuation plus LLM refinement
Speech act aggregation	Rolling-window LLM CONTINUE/NEW decisions
Politeness framework	Brown and Levinson (1987) four superstrategies
Scoring method	Likert 0–5 multi-label, independent per strategy
Output formats	Excel (.xlsx) with embedded thumbnails, CSV backup

Note. All inference uses greedy decoding (do_sample = False) to ensure reproducibility. The pipeline is engineered to operate within the 22 GB VRAM ceiling typical of consumer GPU environments accessible to applied linguistics researchers.

Table 2
Corpus Segmentation Statistics

Metric	Value	M (SD)	Mdn
Total duration (seconds)	719.08	—	—
Total word tokens	1,986	—	—
Pause markers (≥ 0.5 s)	99	—	—
Propositional units	160	—	—
Speech-act units	40	—	—
Propositions per speech act	—	4.00 (3.91)	2.0
Words per speech act	—	50.2 (50.5)	29.5
Speech-act duration (s)	—	16.80 (20.80)	7.86
Single-proposition acts	17 (42.5%)	—	—
Multi-proposition acts	23 (57.5%)	—	—

Note. The maximum aggregated speech-act unit contained 16 propositions, representing extended narrative discourse. Aggregation ratio of 25.0% (40/160) reflects the average bundling of four propositions per speech act.

Table 3
Frequency of Dominant Politeness Superstrategy

Superstrategy	n	%	Cumulative %
Bald On-Record	25	62.5	62.5
Positive Politeness	7	17.5	80.0
Off-Record	7	17.5	97.5
Negative Politeness	1	2.5	100.0
Total	40	100.0	—

Note. Dominant strategy is the superstrategy with the highest Likert score (0–5) for a given speech act. In case of ties (n = 2), the model selected the more pragmatically central strategy as indicated in its rationale. The Bald On-Record predominance reflects the assertive declarative register typical of motivational public speaking.

Table 4
Descriptive Statistics of Likert Scores by Superstrategy

Superstrategy	M	SD	Mdn	Range	n (≥ 4)	n (= 0)
Bald On-Record	4.12	1.04	4	0–5	32 (80.0%)	1
Positive Politeness	1.95	1.65	2	0–5	6 (15.0%)	12
Negative Politeness	1.15	1.00	1	0–4	1 (2.5%)	11
Off-Record	2.58	1.22	3	0–5	7 (17.5%)	4

Note. Likert scoring is independent per superstrategy, allowing strategy blending within a single speech act (e.g., simultaneous Bald = 5 and Positive = 3). N = 40 speech acts. Off-Record shows substantial mean presence (M = 2.58) due to the speaker's frequent use of rhetorical questions, metaphors, and contrastive framing devices.

Politeness Strategy Distribution

The pipeline produced complete annotation coverage across all 40 speech acts, with zero inference errors. Table 3 reports the frequency distribution of dominant politeness superstrategies. Bald On-Record emerged as the predominant strategy in 25 speech acts (62.5%), followed by Positive Politeness and Off-Record at equal frequency (7 speech acts each, 17.5%), and Negative Politeness as the rarest (1 speech act, 2.5%). This distribution aligns with theoretical expectations for public-speaking monologue, a genre characterized by an assertive declarative register where face-threatening acts directed at specific hearers are minimal.

The substantial Off-Record presence as dominant strategy reflects the speaker's characteristic rhetorical style. Speech acts dominated by Off-Record typically realize rhetorical questions, metaphorical framing, and ironic contrasts. These constructions invite audience inference rather than asserting propositions directly, consistent with Brown and Levinson's (1987) characterization of Off-Record as a strategy allowing plausible deniability while signaling a position. Negative Politeness scarcity is consistent with the speaker's stylistic avoidance of explicit hedging, apologies, or formal deference markers.

Multi-Label Scoring and Strategy Blending

Multi-label Likert scoring captured dimensions that single-label classification would obscure. Table 4 reports the descriptive statistics for each superstrategy. Bald On-Record showed the highest mean (M = 4.12, SD = 1.04), with 80% of speech acts scoring 4 or higher. This indicates that direct factual assertion was pervasive across most speech acts even when it was not the dominant strategy.

The substantial mean for Off-Record (M = 2.58) and Positive Politeness (M = 1.95) demonstrates that these strategies frequently co-occurred with Bald On-Record rather than appearing only when dominant. Specifically, strategy blending affected half of the corpus: 50.0% of speech acts received scores of 3 or higher on at least two superstrategy dimensions, indicating substantive co-presence of multiple strategies within a single pragmatic unit. This finding empirically validates the multi-label scoring design choice: forcing single-label classification would systematically misrepresent the layered nature of natural politeness work.

Multimodal Output Structure

Figures 1 and 2 illustrate the multimodal annotation output for a sample of speech acts. Figure 1 displays three sampled frames per unit alongside the verbal transcript and the four-superstrategy Likert scores. Figure 2 displays the corresponding rationale and modality-specific descriptions of facial conduct, gesture, and emotion. Inspection indicates that the model produces consistent, formally English-formatted descriptions that are temporally aligned with the verbal transcript and visually coherent with the source frames.

Table 5
Diagnostic Flag Distribution

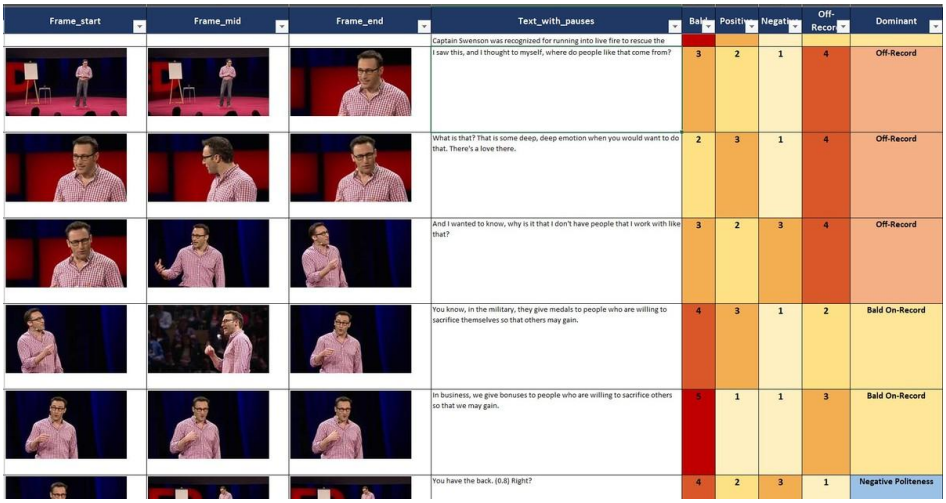
Flag Category	n	%	Action
Visual hallucination	1	2.5	Exclude
Text artifact (broadcast metadata)	1	2.5	Exclude
Short act (< 3 s, < 8 words)	5	12.5	Review
Long act (> 60 s)	2	5.0	Review
Uncertain dominant (tied scores)	2	5.0	Review
Any flag (union)	8	20.0	Inspect
Clean units	32	80.0	Accept

Note. Flags are not mutually exclusive; one unit may trigger multiple flags. Visual hallucination indicates units in which the model described non-existent visual content, typically at broadcast transitions lacking the speaker.

Diagnostic Flag Distribution

Of the 40 annotated speech acts, 32 (80.0%) were clean across all diagnostic categories, while 8 (20.0%) triggered one or more flags. The most common flag categories were short acts (12.5%), uncertain dominant scoring (5.0%), and long acts (5.0%). Visual hallucination and text artifact flags both affected the same single speech act, which contained TED broadcast metadata rather than the speaker's own

speech. The 80% clean rate indicates that VLM-based annotation produces analytically usable output for the substantial majority of speech acts, with a manageable minority requiring expert review.



Frame_start	Frame_mid	Frame_end	Text_with_pauses	Bald	Positive	Negative	Off-Record	Dominant
			Captain Swenson was recognized for running into live fire to rescue the					
			I saw this, and I thought to myself, where do people like that come from?	3	2	1	4	Off-Record
			What is that? That is some deep, deep emotion when you would want to do that. There's a love there.	2	3	1	4	Off-Record
			And I wanted to know, why is it that I don't have people that I work with like that?	3	2	3	4	Off-Record
			You know, in the military, they give medals to people who are willing to sacrifice themselves so that others may gain.	4	3	1	2	Bald On-Record
			In business, we give bonuses to people who are willing to sacrifice others so that we may gain.	5	1	1	3	Bald On-Record
			You have the back. (0.8) Right?	4	2	3	1	Negative Politeness

Figure 1
Multimodal annotation output excerpt.

Note. Excerpt from the pipeline output showing three sampled video frames per speech-act unit, the verbal transcript with pause markers, the four-superstrategy Likert scores (Bald, Positive, Negative, Off-Record) on a 0–5 scale, and the dominant strategy assignment. Six consecutive speech-act units are shown. Color coding of Likert scores ranges from light (low) to dark (high). Dominant strategy color coding follows Brown and Levinson superstrategy categories.

Rationale	Face	Gesture	Emotion
The speaker uses indirect questioning ("where do people like that come from?") to avoid direct judgment, supported by ambiguous phrasing and a smiling, non-confrontational expression. His clasped hands and relaxed stance suggest deference, while the rhetorical question invites audience inference, making Off-Record dominant. The mild smile and lack of direct address indicate minimal negative politeness, and the question's vagueness supports Off-Record scoring.	Neutral expression, no direct eye contact or facial tension; slight smile in middle frame, suggesting mild engagement.	Hands clasped and slightly open, minimal movement; posture relaxed, no emphatic gestures.	Slight smile (cues: middle frame close-up, lips upturned, eyes slightly crinkled). BALD_ON_RECORD: 3 POSITIVE_POLITENESS: 2 NEGATIVE_POLITENESS: 1 OFF_RECORD: 4
The speaker uses vague, metaphorical phrasing ("deep, deep emotion") and rhetorical framing ("what is that?") to invite interpretation, supported by open gestures and warm expression. He avoids direct commands or imposition, aligning with Off-Record. The "love there" is understated, not explicit, reinforcing ambiguity. Facial warmth and hand gestures suggest inclusion, but the core strategy is indirectness to preserve plausible deniability.	Neutral expression, slight smile, no direct eye contact or facial tension; no overt face-threatening acts.	Hands open and gesturing mid-speech, palms up; slight head tilt in middle frame, indicating openness.	Slight smile (cues: mouth curve, eyes slightly crinkled in first frame); engaged, warm tone. BALD_ON_RECORD: 2 POSITIVE_POLITENESS: 3 NEGATIVE_POLITENESS: 1 OFF_RECORD: 4
The speaker's question "why is it that I don't have people that I work with like that?" is indirect, inviting inference (Off-Record 4), while his open gestures and neutral face suggest he's not imposing (Negative Politeness 3) and avoids direct confrontation. The question's phrasing implies a personal longing, not a demand, aligning with Positive Politeness 2. The ambiguity and rhetorical nature of the question dominate the speech act.	Neutral expression, slight head tilt, no direct eye contact with audience; no visible facial tension or emotional display.	Hands open and gesturing outward in first frame, then slightly inward in second, then closed fist in third — indicating reflection or emphasis.	Neutral (cues: slight head tilt, open mouth mid-speech, no strong facial expression or gesture) BALD_ON_RECORD: 3 POSITIVE_POLITENESS: 2 NEGATIVE_POLITENESS: 3 OFF_RECORD: 4
The speaker uses direct declarative language ("they give medals to people who are willing to sacrifice") without hedging or indirectness, matching Bald On-Record. His enthusiastic expression and hand gestures reinforce directness. Positive Politeness is present via "you know" implying shared experience, but is secondary. No negative politeness markers or off-record strategies are evident.	Neutral expression, slight smile, no direct eye contact with audience; no visible facial tension or avoidance.	Hand gestures consistent with explanation; fingers slightly curled, palm facing inward, indicating emphasis on shared understanding.	Enthusiastic (cues: open mouth, raised eyebrows, slight smile, engaged posture) BALD_ON_RECORD: 4 POSITIVE_POLITENESS: 3 NEGATIVE_POLITENESS: 1 OFF_RECORD: 2
The speaker directly states a contrast without hedging or indirect phrasing, matching Bald On-Record (5). The thumbs-up gesture and slight smile suggest positive framing, but no in-group solidarity or deference. Off-record (3) is present via implied rhetorical contrast, inviting audience inference. No negative face strategies or positive politeness markers are evident.	Neutral expression, no direct eye contact or facial tension; slight head tilt suggests mild engagement.	Hand gestures are minimal; right hand points slightly, then gives thumbs-up, indicating emphasis and approval.	Neutral to slightly positive (cues: slight smile, raised eyebrows, thumbs-up gesture). BALD_ON_RECORD: 5 POSITIVE_POLITENESS: 1 NEGATIVE_POLITENESS: 1 OFF_RECORD: 3

Figure 2
Qualitative annotation layer excerpt.

Note. Excerpt from the pipeline output showing the model-generated rationale, facial conduct description, bodily conduct description, and emotion inference. Each row corresponds to one speech-act unit and integrates verbal phrases from the transcript with observations of facial and bodily conduct across the three sampled frames. The rationale column documents the model's reasoning for the assigned Likert scores.

Implications for Downstream Coding

The structural consistency of the pipeline output supports its use as input to downstream rubric-based coding workflows. Because the politeness scores and qualitative descriptions are generated in a predictable format and aligned with verbal and temporal anchors, expert annotators can apply rubrics directly to the model output rather than re-coding source video from scratch. This division of labor positions the vision-language model as a first-pass describer that reduces the cognitive load on human annotators, who can then focus on theoretically motivated functional validation rather than raw observational description. Such a workflow has been argued for in recent LLM-assisted annotation studies in discourse and pragmatics (Ostyakova et al., 2025; Yu et al., 2024).

Two methodological observations follow from the present results. First, the pipeline produces output that is sufficient for first-pass annotation but does not yet substitute for expert validation. The descriptions and scores are formally consistent and visually grounded, but they have not been validated against expert pragmatic judgment through inter-rater reliability statistics. Second, the pipeline's modular design permits substitution. The Qwen3-VL model can be replaced by alternative VL LLMs without architectural change; the four-superstrategy framework can be replaced by alternative pragmatic taxonomies; the speech-act unit can be replaced by alternative discourse units. Future work can therefore benchmark alternative configurations within the same pipeline.

Conclusion

This study has presented a proof of concept for a computational pipeline that integrates speech recognition, prosodic analysis, vision-language modeling, and structured politeness annotation to produce multimodal descriptions of pragmatic markers. The pipeline operates within the 22 GB VRAM constraint typical of consumer GPU environments and produces formally consistent output across all units of analysis. Applied to a twelve-minute public-speaking sample, the pipeline yielded 40 fully annotated speech acts with theoretically coherent strategy distributions and multi-label scoring that captured strategy blending across half of the corpus. The contribution of the work is methodological: the pipeline is demonstrated to be technically feasible and structurally tractable for multimodal pragmatic annotation.

Three limitations require acknowledgment. First, the proof-of-concept design provides no inter-rater validation or accuracy benchmarking against expert coding. The pipeline output may exhibit systematic biases that only expert comparison can reveal. Second, the application is limited to a single public-speaking sample. Generalization to other genres, languages, and interactional configurations remains untested. Third, the pipeline operates through prompt engineering on a general-purpose foundation model rather than through fine-tuning on pragmatic annotation data. Fine-tuning is likely to improve domain-specific accuracy.

Three directions for future work follow. First, expert validation should be conducted on the present pipeline output against human coders applying established politeness annotation schemes, with inter-rater reliability computed using appropriate ordinal-agreement statistics. Second, the pipeline should be extended to corpora spanning multiple speakers, genres, and languages, including Indonesian public-speaking data to enable cross-cultural comparison of multimodal pragmatic markers. Third, the pipeline output can serve as a foundation for parameter-efficient fine-tuning experiments using LoRA adapters on Qwen3-VL, potentially yielding substantial improvement over the present zero-shot baseline.

References

Aijmer, K. (2013). *Understanding pragmatic markers: A variational pragmatic approach*. Edinburgh University Press.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., … Lin, J. (2025). Qwen2.5-VL technical report. *arXiv*. https://doi.org/10.48550/arXiv.2502.13923

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press. https://doi.org/10.1017/CBO9780511813085

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge. https://doi.org/10.4324/9780203970034

Lin, F. (2025). Vision language models: A survey of 26K papers. *arXiv*. https://doi.org/10.48550/arXiv.2510.09586

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

Ostyakova, L., Mikhailova, A., & Konovalov, V. (2025). Redefining annotation practices: Leveraging large language models for discourse annotation. In A. Panchenko et al. (Eds.), *Analysis of images, social networks and texts (AIST 2024)* (Lecture Notes in Computer Science, Vol. 15419). Springer. https://doi.org/10.1007/978-3-031-88036-0_7

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*. https://doi.org/10.48550/arXiv.2212.04356

Schiffrin, D. (1987). *Discourse markers*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611841

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apologies. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.23087.yu