A Reproducible Multi-Architecture Baseline for Token-Level Chinese Metaphor Identification under the MIPVU Framework

Anonymous ACL submission

Abstract

Metaphor is pervasive in everyday language, yet token-level identification of metaphor-related words in Chinese under the MIPVU framework remains under-explored relative to English. We present a reproducible multi-architecture baseline for token-level metaphor identification on the PSU Chinese Metaphor Corpus (PSU CMC), comparing three model families: (i) encoder fine-tuning with Chinese RoBERTa-wwm-ext-large; (ii) MelBERT adapted to Chinese using a newly constructed basic-meaning resource derived from the Modern Chinese Dictionary, 7th edition (MCD7, 74,823 entries with 71.51% PSU CMC vocabulary coverage); and (iii) Qwen3.5-9B fine-tuned with QLoRA as an instruction-tuned generative baseline. Across five fixed seeds, MelBERT MIP-only achieves the strongest performance at 0.7281 ± 0.0050 test positive F1, marginally above the full three-channel MelBERT and clearly above plain RoBERTa, while the Qwen QLoRA configuration trails encoder baselines by approximately 11 F1 points, with the gap concentrated in recall. Our findings suggest that the SPV channel of MelBERT contributes little reliable signal for predominantly conventional Chinese metaphor, and that the precision-favoring asymmetry of generative LLMs invites a hybrid encoder-filter / LLM-booster design. Split manifests, per-seed outputs, the MCD7 basic-meaning embedding pipeline, and training scripts will be released with the camera-ready version.

1 Introduction

Metaphor is pervasive in everyday language (Steen et al., 2010), and its computational identification is a long-standing problem with applications in sentiment analysis, machine translation, discourse understanding, and language education. The Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) and its operationalization MIPVU (Steen et al., 2010) provide the dominant linguistic protocol for token-level metaphor annotation: each lexical unit is judged metaphor-related if its contextual meaning differs from, but can be understood by comparison with, a more concrete or basic meaning.

For English, MIPVU has driven a substantial research program. Successive systems—from biLSTM-CRF taggers to Transformer-based sequence labelers—have been evaluated on the VU Amsterdam Metaphor Corpus and TOEFL essays (Mao et al., 2019; Su et al., 2020; Gong et al., 2020), and the most influential recent architecture, MelBERT (Choi et al., 2021), explicitly grounds its design in MIP and Selectional Preference Violation theory. For Chinese, by contrast, the picture is markedly thinner. Lu and Wang (2017) introduced the PSU Chinese Metaphor Corpus (PSU CMC), the only widely available token-level MIPVU-annotated Chinese corpus, and yet PSU CMC has remained an under-evaluated benchmark: modern encoder fine-tuning, MelBERT-style lexical fusion, and instruction-tuned LLMs have not been systematically compared on it under a shared seed protocol.

This paper addresses that gap. We treat the question pragmatically: *what does a careful, reproducible baseline for token-level metaphor identification on PSU CMC look like in 2026?* Concretely, we compare three model families that span the methodological space: (i) standard encoder fine-tuning with Chinese RoBERTa-wwm-ext-large (Cui et al., 2021); (ii) MelBERT adapted to Chinese, requiring a Chinese basic-meaning resource that does not yet exist in open form; and (iii) Qwen3.5-9B fine-tuned with QLoRA (Dettmers et al., 2023; Hu et al., 2022) as an instruction-tuned generative baseline.

The MelBERT adaptation requires resolving a resource gap. The original MelBERT relies on WordNet first-sense glosses to supply each token's basic meaning. No analogous open Chinese resource

exists. We therefore construct a basic-meaning resource derived from the Modern Chinese Dictionary, 7th edition (MCD7), comprising 74,823 entries with parseable basic meanings encoded as 1024-dimensional vectors, with 71.51% coverage of PSU CMC vocabulary. This resource is itself a contribution of the paper; the derived artifacts (embeddings, mappings, pipeline) will be licensed for research use, while the original copyrighted gloss text is not redistributed.

Across five fixed seeds, the strongest configuration we evaluate is MelBERT MIP-only (i.e., MelBERT with the SPV channel removed), at 0.7281 ± 0.0050 test positive F1. This is marginally above the full three-channel MelBERT (0.7270 ± 0.0069) and clearly above plain RoBERTa fine-tuning (0.7142 ± 0.0121). Qwen3.5-9B with QLoRA trails the encoder-based baselines by roughly 11 absolute F1 points (0.6157 ± 0.0113), with the gap concentrated in recall and amplified on the fiction register. We discuss the underlying causes in Section 5.

Our contributions are threefold: a reproducible multi-architecture baseline on PSU CMC with five-seed runs across four configurations; a Chinese MIP basic-meaning resource derived from MCD7, enabling MelBERT-style lexical fusion in Chinese for the first time; and empirical findings on Chinese metaphor identification, including the unexpectedly competitive MelBERT MIP-only ablation and the precision-favoring asymmetry of QLoRA-adapted generative LLMs.

2 Related Work

We situate our work along three axes: the MIPVU annotation framework and its Chinese adaptation (§2.1), computational methods for metaphor identification (§2.2), and lexical resources for basic-meaning representation (§2.3).

2.1 MIPVU and Token-level Metaphor Resources

The Metaphor Identification Procedure (MIP) was introduced by the Pragglejaz Group (Pragglejaz Group, 2007) as a reproducible inter-annotator protocol for tagging metaphor-related words in running text. Steen et al. (2010) extended it into MIPVU, which adds explicit treatment of indirect metaphor, direct metaphor, and borderline cases, and was applied to construct the VU Amsterdam Metaphor Corpus, the canonical English MIPVU-annotated benchmark. For Chinese, Lu and Wang (2017) adapted MIPVU to Mandarin and constructed the PSU Chinese Metaphor Corpus (PSU CMC) by sampling documents from the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004); PSU CMC is the corpus we use in this paper.

Other Chinese metaphor corpora have been released, but they target different tasks and annotation schemes. CMC (Li et al., 2023) provides sentence-level metaphor labels with a heavy positive-class skew (∼91% positive), making it a metaphor-rich classification benchmark rather than a representative running-text identification benchmark. CMDAG (Shao et al., 2024) annotates metaphor with grounds (rationales) for metaphor generation. Neither aligns directly with MIPVU's token-level identification task on naturally distributed text, so we do not evaluate cross-corpus transfer in this paper.

Wang et al. (2019) adapted the MIPVU protocol specifically for Chinese in a chapter of the multi-language MIPVU volume (Wang et al., 2019; Nacey et al., 2019), documenting challenges including word segmentation ambiguity, grammaticalized prepositions, and compound-internal metaphor. Their adapted protocol defines Chinese-specific metaphor flag (MFlag) words, including pre-source and post-source markers that signal direct metaphor but are themselves not annotated as metaphor-related in PSU CMC.

2.2 Metaphor Identification Methods

Computational approaches to token-level metaphor identification have evolved through three overlapping waves. Sequence-tagging neural models, beginning with biLSTM-CRF and ELMo-based architectures (Mao et al., 2019), treat metaphor identification as a standard tagging task and were the dominant paradigm prior to the widespread use of pretrained Transformer encoders. The introduction of BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) reshaped the field: DeepMet (Su et al., 2020) won the 2020 VUA Metaphor Detection Shared Task using a reading-comprehension formulation over RoBERTa with linguistic features; IlliniMet (Gong et al., 2020) combined RoBERTa with WordNet, VerbNet, POS, and concreteness features. MelBERT (Choi et al., 2021) departed from this "RoBERTa-plus-features" paradigm by encoding two metaphor-theoretic inductive biases (MIP and SPV) directly into its architecture. MelBERT remains the strongest published

BERT-family approach we adapt; its dependence on a basic-meaning resource is what motivates our MCD7 construction.

Instruction-tuned large language models remain less settled for token-level MIPVU identification because generation must commit to discrete metaphor labels without a calibrated probability head. Our Qwen3.5-9B with QLoRA experiments test this paradigm in Chinese against MIPVU encoder baselines.

For Chinese specifically, Zhang et al. (2021) proposed SaGE, a syntax-aware GCN with ELECTRA model achieving 85.22% macro-F1 on the CCL2018 Chinese metaphor evaluation dataset (Zhang et al., 2021). However, CCL2018 is a sentence-level three-class task (verb metaphor / noun metaphor / literal), fundamentally different from PSU CMC's token-level binary identification under MIPVU. No prior work has reported encoder fine-tuning, MelBERT-style lexical fusion, or systematic LLM comparison on PSU CMC.

In the LLM prompting paradigm, Huang and Liu (2026) reported a GPT-4-based, interpretable MIPVU rule-script framework on PSU CMC in an arXiv preprint (Huang and Liu, 2026). Fuoli et al. (2025) systematically compared prompt engineering, retrieval-augmented generation (RAG), and fine-tuning for English metaphor identification, and reported strongest performance for the fine-tuned setting (Fuoli et al., 2025). Our work extends this prompt-versus-fine-tune comparison to the Chinese MIPVU setting, where we additionally include encoder baselines that the LLM literature has not compared against.

2.3 Lexical Resources for Basic Meaning

MelBERT-style lexical fusion requires per-token basic-meaning representations. For English, this role is filled by WordNet first-sense glosses, which are open-licensed and computationally accessible. For Chinese, the lexical-resource situation is heterogeneous: HowNet, Chinese WordNet, and BCC have been used for various Chinese NLP tasks, but each has limitations as a MelBERT substrate—HowNet's sememe representation does not align with MIPVU's notion of basic meaning, Chinese WordNet has sparser coverage than its English counterpart, and BCC is a corpus rather than a lexicon. We therefore construct a basic-meaning resource directly from MCD7, the most widely cited authoritative reference dictionary of modern Chinese, with full extraction and encoding pipelines to be released after review.

3 Method

This section describes the evaluation corpus and its split protocol (§3.1–3.2), the basic-meaning resource we construct for MelBERT adaptation (§3.3), the three model configurations under comparison (§3.4), and the shared training and evaluation protocol (§3.5).

3.1 PSU Chinese Metaphor Corpus

The PSU Chinese Metaphor Corpus (PSU CMC) (Lu and Wang, 2017) is a multi-register Chinese corpus annotated with token-level metaphor labels following the Metaphor Identification Procedure VU (MIPVU) (Steen et al., 2010). The text is sampled from the Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery and Xiao, 2004), a one-million-word balanced corpus of written Mandarin, from which Lu and Wang (2017) drew 75 documents covering three registers: academic prose, fiction, and news.

Each lexical unit is assigned a binary metaphor flag based on the contextual-vs-basic meaning contrast central to MIPVU. As shown in Table 1, the corpus contains 1,724 sentences and 35,746 tokens, of which 3,272 tokens (9.16%) are labeled as metaphor-related words. Metaphor density varies markedly across registers, ranging from 6.36% in news to 13.67% in academic.

3.2 Data Split

We adopt a *file-level* 70/10/20 train/dev/test split with seed 42, ensuring that no document appears in more than one partition. Sentence-level splits, common in earlier metaphor identification work, can leak document-level cues since adjacent sentences from the same source share register, topic, and stylistic conventions, leading to optimistic estimates of generalization.

The resulting split contains 1,182 train, 198 dev, and 344 test sentences (from 52, 8, and 15 documents respectively; see Table 1). The split manifest—i.e., the document IDs assigned to each partition—will be released as part of our reproducibility package after review, and all experiments throughout this paper use this split without modification.

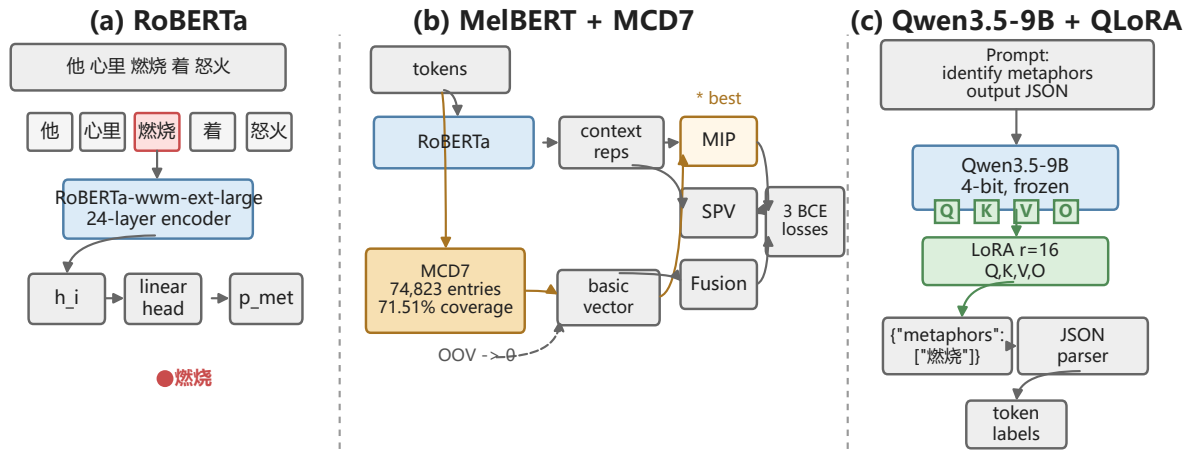

Figure 1: Architectures of the three model families compared in this work. (a) RoBERTa-wwm-ext-large with a linear token-classification head. (b) MelBERT extends the same encoder with two metaphor-theoretic channels (MIP and SPV) plus a fusion head; the MIP channel consumes basic-meaning vectors looked up from our MCD7-derived resource (74,823 entries, 71.51% PSU CMC coverage), with a zero-vector fallback for OOV tokens. (c) Qwen3.5-9B with QLoRA adapters on the four attention projection matrices, framed as JSON-generation followed by deterministic parsing. The example sentence *ta xinli ranshao zhe nuhuo* ('anger burned in his heart') contains a conventional metaphor on *ranshao* ('burn').

Table 1: PSU CMC dataset statistics by register and split.

Subset	#Docs	#Sentences	#Tokens	#Metaphor	%Metaphor
Academic	30	487	11735	1604	13.67
News	25	528	12027	765	6.36
Fiction	20	709	11984	903	7.54
Total	75	1724	35746	3272	9.15
Train	52	1182	24887	2254	9.06
Dev	8	198	4227	356	8.42
Test	15	344	6632	662	9.98

3.3 Modern Chinese Dictionary Basic-Meaning Resource

A central component of the MelBERT architecture (Section 3.4.2) is the *basic meaning* of each lexical unit—the most concrete or primary sense, used as the contrastive anchor in the MIP channel. The original MelBERT formulation for English uses WordNet first-sense glosses (Choi et al., 2021). For Chinese, no equivalent open lexical resource exists. We therefore construct one from the *Modern Chinese Dictionary*, 7th edition (Dictionary Editorial Office, Institute of Linguistics, Chinese Academy of Social Sciences, 2016), hereafter MCD7—the authoritative Chinese-language reference dictionary maintained by the Institute of Linguistics, Chinese Academy of Social Sciences, and published by the Commercial Press.

3.3.1 Construction Pipeline

We derive the resource from MCD7 in five stages: (1) format extraction from the source MDX, yielding 74,823 raw entries; (2) sense parsing of each entry's definitions; (3) basic-meaning selection, taking the most concrete or primary sense per MIPVU principles, with a fallback to the full headword definition for unparseable entries; (4) cross-reference resolution, recursively following 'see' / 'same as' redirects with circular-reference detection; and (5) embedding encoding, mapping each basic-meaning text to a 1024-dimensional vector via the Chinese RoBERTa-wwm-ext-large `[CLS]` representation. Detailed cross-reference resolution statistics are reported in Appendix A. The result is a one-to-one mapping from dictionary headwords to basic-meaning vectors, consumed directly by the MelBERT MIP channel (Section 3.4.2).

3.3.2 Statistics and Coverage

The dictionary contains 74,823 entries; 99.33% have parseable basic-meaning text (the remaining 0.67% retain only the headword as fallback). Of these entries, 60,075 (80.29%) are single-sense and 14,748 (19.71%) are multi-sense (10,859 with two senses, 3,889 with three or more), with a mean of 1.31 and a maximum of 24 senses per entry. Coverage of the PSU CMC vocabulary is 71.51%—that is, 5,094 of 7,123 unique tokens in PSU CMC have a corresponding dictionary entry. The remaining 28.49% of PSU CMC tokens fall outside the dictionary, primarily proper nouns, register-specific compounds, and rare expressions. These tokens receive a zero-vector fallback at MelBERT input, marked via an out-of-vocabulary mask.

3.3.3 Use in MelBERT and Limitations

For each input token, MelBERT looks up its dictionary entry and uses the encoded basic-meaning vector as the input to its MIP channel. We use *only the first sense* per entry, following the original MelBERT design. This means the 19.71% of multi-sense entries are underutilized; richer integration of the full sense inventory is a natural extension we discuss in Section 5.

3.3.4 License and Release

The Modern Chinese Dictionary is copyrighted by the Commercial Press; we therefore do not redistribute the original gloss text. After review, we will release: (a) the per-entry 1024-dimensional embedding vectors derived from the basic-meaning text; (b) entry token-to-index mappings; and (c) the full extraction and encoding pipeline as scripts. Users wishing to reproduce or extend the resource must obtain the dictionary independently. Artifacts (a)–(c) will be distributed under MIT license (code) and CC BY 4.0 (derived data) for research use.

3.4 Model Architectures

Figure 1 summarizes the three architectures. We compare three model families that span the current methodological space for token-level metaphor identification.

3.4.1 RoBERTa-wwm-ext-large (Encoder Fine-tuning)

Our first baseline is standard token-classification fine-tuning of Chinese RoBERTa-wwm-ext-large (Cui et al., 2021), a 24-layer Transformer encoder with whole-word-masking pre-training on Chinese Wikipedia and EXT data. We add a linear classification head on top of each token's final hidden state, producing a binary metaphor probability per token. Training uses standard cross-entropy loss; no auxiliary signals or external resources are involved.

3.4.2 MelBERT

Architecture. MelBERT (Choi et al., 2021) extends an encoder backbone with two parallel channels grounded in metaphor identification theory: (1) the Metaphor Identification Procedure (MIP) channel contrasts each token's contextualized representation against its basic-meaning vector (Section 3.3); (2) the Selectional Preference Violation (SPV) channel contrasts the token's local context with the token's contextualized representation, flagging selectional anomalies. Both channels feed into independent classification heads, and a third *fusion head* combines them. The training loss is a weighted sum of three binary cross-entropy losses (one per head).

Chinese adaptation. We use the same Chinese RoBERTa-wwm-ext-large backbone as in Section 3.4.1 and replace the WordNet-based basic-meaning embeddings of the original English MelBERT with our MCD7-derived embeddings (Section 3.3). Tokens not in MCD7 receive a zero-vector fallback marked via an out-of-vocabulary mask.

Channel ablations. In addition to the standard three-channel *Full* configuration, we evaluate two ablations: *MIP-only* (SPV channel and fusion head removed; loss is single-channel binary cross-entropy) and *SPV-only* (mirror ablation).

3.4.3 Qwen3.5-9B with QLoRA (Instruction-tuned Generation)

Backbone. Our third model is Qwen3.5-9B, a 9-billion-parameter instruction-tuned language model. Unlike RoBERTa and MelBERT, Qwen treats metaphor identification as a generation task: given a sentence as prompt, the model generates a structured output identifying metaphor positions, which we then parse deterministically.

Parameter-efficient adaptation. We use QLoRA (Dettmers et al., 2023), which combines 4-bit NF4 quantization of the frozen backbone with low-rank adapter training (Hu et al., 2022). Adapters are inserted into the four attention projection matrices (query, key, value, output)

with rank $r = 16$ and scaling factor $\alpha = 32$. This reduces trainable parameters to under 1% of the full model while keeping the backbone in 4-bit precision (peak GPU memory: 15.3 GB on a single RTX 5090).

Task formulations. The choice of how to encode token-level metaphor identification as a generation task is non-trivial. We systematically compare six task formulations (Q1–Q8 in Table 3), spanning classification-style and generation-style designs:

- **Classification-style**: token-level head over hidden states (Q1), and BIO-tagged span prediction (Q4).
- **Generation-style**: free-form JSON listing metaphor tokens (Q2), QA-style natural language (Q6), and structured generation with token-id constraints (Q8 v1/v2).

Q2 (Generative JSON Extraction) emerged as the strongest formulation in single-seed experiments (Table 3); we adopt it for the 5-seed main comparison reported in Section 4.1.

3.5 Training and Evaluation Protocol

All experiments run on a single NVIDIA RTX 5090 (32 GB VRAM), with Qwen3.5-9B loaded in 4-bit NF4 quantization and bf16 compute.

For each main model (RoBERTa, MelBERT Full, MelBERT MIP-only, and Qwen Q2), we run training and evaluation across five fixed seeds: {42, 123, 2024, 7, 31415}. The seed list is fixed in advance and released alongside our code after review, so all multi-seed numbers in this paper correspond to the same seed set across architectures. We report mean±std with population standard deviation (ddof = 0) throughout, matching the convention used in our internal aggregation scripts.

Table 4 summarizes the final training configurations. RoBERTa and MelBERT both use the Chinese RoBERTa-wwm-ext-large backbone with learning rate 5×10^{-5}, effective batch size 16, up to 10 epochs, and maximum sequence length 256. Qwen Q2 uses learning rate 2×10^{-4}, effective batch size 16 (batch 1 with gradient accumulation 16), up to 3 epochs, and maximum sequence length 1024 to accommodate the prompt-plus-response generative format. The Qwen configuration is the one verified against per-seed `train_summary.json` files; an early planning document contained inaccurate values, which we do not adopt.

For all models, training uses early stopping with patience 3 (for RoBERTa and MelBERT) or patience 2 (for Qwen) on dev-set monitoring. RoBERTa and MelBERT monitor dev positive-class F1 directly; Qwen Q2 monitors dev cross-entropy loss, since per-batch dev F1 was not logged during training.

We report *positive-class F1* (denoted *Test pos-F1*) as the primary metric. PSU CMC is heavily class-imbalanced (9.16% positive at the token level), so macro F1 is dominated by the trivial negative-class score and inflates absolute numbers without reflecting metaphor identification quality. We report macro F1 alongside positive F1 for completeness, but all comparisons in this paper, including the register-level breakdowns, use positive-class F1.

For the Qwen generative configuration, predictions are obtained by deterministic JSON parsing of the model's output. Tokens listed in the parsed JSON receive a positive label; all others receive a negative label. Failures of JSON parsing (e.g., malformed brackets) are counted as a separate *parse failure rate*; in practice this rate is below 0.3% across all 5 Qwen seeds and does not materially affect F1.

Per-register F1 is computed by partitioning the test set into the three registers (academic, news, fiction) and computing positive-class F1 within each subset. Register subsets share the same gold labels and are non-overlapping. All metric aggregation uses ddof = 0 across seeds.

4 Experiments

We first report the main cross-architecture comparison (§4.1), per-register performance (§4.2), and Qwen task-form comparison (§4.3). We discuss the Qwen precision–recall asymmetry in Appendix C.

4.1 Main Comparison

Table 2 reports five-seed test performance for RoBERTa, MelBERT (Full and MIP-only), and Qwen3.5-9B with QLoRA. MelBERT MIP-only is the strongest and most stable configuration, while Qwen Q2 trails the encoder baselines chiefly through lower recall despite comparable precision. The Full–MIP-only gap is within one standard deviation, so we do not claim statistical superiority for MIP-only; we instead treat the consistency of the pattern across seeds and metrics as evidence that the SPV channel does not contribute reliably positive signal in our setting.

The competitive performance of MIP-only mer-

Table 2: Test set performance of four model configurations on PSU CMC, reported as mean ± standard deviation over 5 seeds (population std, ddof = 0).

Model	Test pos-F1	Macro F1	Precision	Recall
RoBERTa-large	0.7142 ± 0.0121	0.8421 ± 0.0062	0.7536 ± 0.0206	0.6807 ± 0.0367
MelBERT (full)	0.7270 ± 0.0069	0.8490 ± 0.0036	0.7568 ± 0.0158	0.7003 ± 0.0211
MelBERT (MIP-only)	$\mathbf{0.7281 \pm 0.0050}$	$\mathbf{0.8496 \pm 0.0026}$	$\mathbf{0.7572 \pm 0.0193}$	$\mathbf{0.7021 \pm 0.0205}$
Qwen3.5-9B (Q2)	0.6157 ± 0.0113	0.7889 ± 0.0057	0.6963 ± 0.0103	0.5526 ± 0.0235

its further examination. The original MelBERT paper (Choi et al., 2021) motivates the SPV channel as complementary to MIP, but on PSU CMC we observe the opposite trend from English VUA: MIP-only matches or slightly exceeds the Full configuration on every aggregate metric (Table 2), with notably lower seed variance.

We discuss two non-exclusive explanations for this pattern in Section 5.

4.2 Per-Register Breakdown

Figure 2 visualizes positive F1 separately for the three registers in PSU CMC. Across all four configurations, academic prose is the easiest register, followed by news, with fiction the hardest. The ordering is consistent with metaphor density: academic prose has the highest metaphor density (13.67%, see Table 1) while fiction has only 7.54% and features more novel, context-dependent metaphor.

The cross-architecture pattern is preserved within each register: MelBERT MIP-only and Full are essentially tied on academic (0.7532 ± 0.0048 vs. 0.7538 ± 0.0084) and news (0.7357 ± 0.0143 vs. 0.7327 ± 0.0068), with MIP-only modestly ahead on fiction (0.6725 ± 0.0271 vs. 0.6694 ± 0.0237). Notably, fiction is also where seed variance is largest for every model (Qwen reaches $\sigma = 0.0378$ on fiction versus 0.0055 on academic), and where Qwen's gap to the encoders is most severe (roughly -0.13 absolute pos-F1 versus MelBERT MIP-only on fiction, versus -0.09 on academic).

Fiction's lower performance and higher variance reflect its lower metaphor density (7.54%, Table 1) and its more novel, context-dependent metaphors compared to conventionalized academic prose. For Qwen specifically, seed 7 collapses on fiction (0.4702 vs. a non-seed-7 mean of 0.5600), producing the outlier visible in Figure 3.

4.3 Qwen Task-Form Comparison

Before the five-seed comparison, we screened six task formulations for Qwen3.5-9B in single-seed experiments (see Appendix B). Generative JSON extraction (Q2) was the strongest at 0.6275 test positive F1, and we adopt it for the main comparison; no formulation reached the encoder baseline. The two weakest formulations fail for format-design rather than model-capacity reasons: BIO span tagging (Q4) collapses on Chinese metaphor's predominantly single-token spans, and structured generation under a 256-token limit (Q8 v1) collapses from supervision-token truncation, recovering once the limit is raised. Full per-formulation results and the failure analysis are given in Appendix B.

5 Discussion

5.1 Why MIP-only matches Full MelBERT

The three-channel ablation gives Full 0.7270 ± 0.0069, MIP-only 0.7281 ± 0.0050, and SPV-only 0.7206 ± 0.0064: the standalone SPV channel is the weakest, while MIP alone matches or slightly exceeds the fused configuration with notably lower seed variance (-28%). We offer two non-exclusive explanations.

First, **conventional metaphor dominance.** A substantial fraction of metaphor in Chinese running text is conventional rather than novel: the metaphorical sense is well-established in the lexicon. For such cases the contextual-vs-basic-meaning contrast (MIP) remains informative, but the metaphor-context pairing is too frequent to look anomalous to SPV. Second, **optimization noise from the three-way loss.** Reducing three heads to one halves competing objectives, and the observed variance reduction is consistent with a less-noisy optimization landscape independent of any linguistic claim about SPV's signal value.

5.2 The Qwen–encoder gap is a recall problem

Qwen Q2 trails the encoder baselines by approximately 11 absolute F1 points, but the gap is concentrated in recall (0.5526 ± 0.0235) rather than precision (0.6963 ± 0.0103). Two structural causes are likely. Generative output forces a discrete

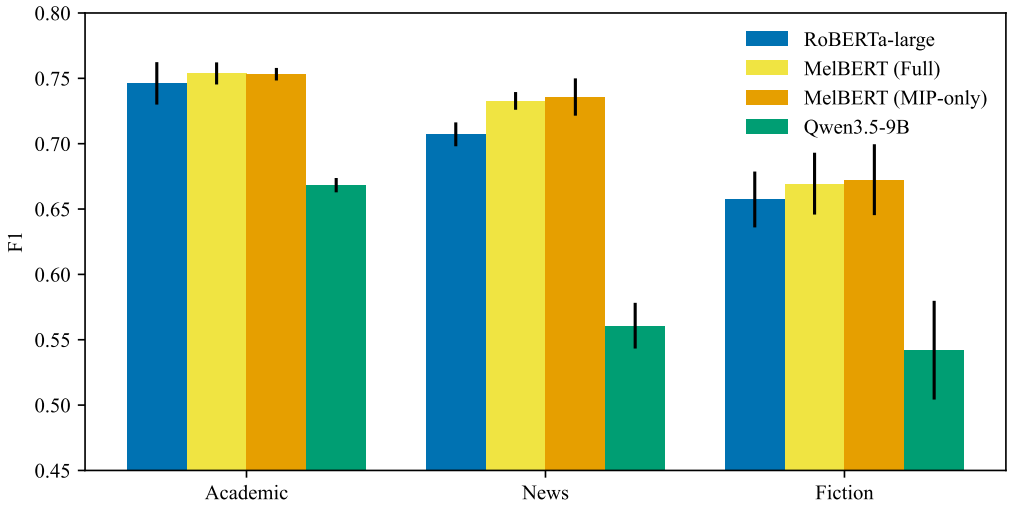

Figure 2: Per-register positive F1 across the four model configurations on PSU CMC test set. Bars show 5-seed means; error bars represent $\pm 1\sigma$ (population std). Registers ordered by MelBERT (MIP-only) F1, descending.

commitment per token with no calibrated probability gradient: tokens whose metaphor status is genuinely ambiguous are dropped rather than included with low confidence. QLoRA further modifies only the four attention projection matrices and may underfit a fine-grained token-level signal compared to a dense classification head trained on full-rank hidden states. The combination—hard discrete decision plus low-rank adaptation—biases Qwen toward the precision-favoring regime we observe. This pattern suggests a natural hybrid: a high-precision encoder filter paired with an LLM recall booster, which we leave to future work.

5.3 Reproducibility and artifacts

For review, artifact links are anonymized. After review, we will release: (a) the file-level train/dev/test split manifest (Section 3.2); (b) per-seed checkpoints, evaluation outputs, and aggregated metrics for all four model configurations; (c) the MCD7 dictionary basic-meaning embedding pipeline and the derived embeddings under the licensing terms in Section 3.3.4; and (d) all training and aggregation scripts. We expect future systems to compare against the configurations and seed protocol reported here, rather than against informal estimates from the literature.

6 Conclusion

We presented a reproducible multi-architecture baseline for token-level Chinese metaphor identification on PSU CMC, comparing encoder fine-tuning, MelBERT with a newly constructed MCD7 basic-meaning resource, and Qwen3.5-9B with QLoRA. MelBERT MIP-only is the strongest configuration at 0.7281 ± 0.0050 test positive F1. Natural extensions include multi-sense MelBERT integration exploiting the 19.71% multi-sense MCD7 entries, metaphor-aware domain-adaptive pre-training as a shared ceiling for the encoder baselines, and the encoder-filter plus LLM-booster hybrid suggested by the precision–recall asymmetry of Section 5. We hope these baselines, configurations, and review-anonymized artifacts provide a useful reference point for future work on Chinese metaphor identification.

Limitations

Several limitations are worth flagging explicitly. First, our MelBERT implementation uses only the first sense per dictionary entry (Section 3.3.3). The 19.71% of multi-sense entries in MCD7 are underutilized; richer sense-aware integration, such as attention-weighted multi-sense aggregation, is a natural follow-up. Second, our Qwen Q2 configuration was selected by single-seed task-form comparison (Section 4.3); a more thorough hyperparameter search across the better-performing formulations could narrow the gap to encoder baselines, though we do not expect it to close. Third, dev-set early-stopping for Qwen used cross-entropy loss rather than dev positive F1 (Section 3.5), a logging artifact rather than a methodological choice; future runs should monitor dev positive F1 directly. Fourth, we evaluate only on PSU CMC test; cross-corpus generalization, for example to CMC (Li et al., 2023) or CMDAG (Shao et al., 2024), is left for future work, particularly because those corpora use different annotation schemes and are not directly comparable.

References

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1763–1773.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186. Association for Computational Linguistics.

Dictionary Editorial Office, Institute of Linguistics, Chinese Academy of Social Sciences. 2016. *Xiandai Hanyu Cidian (Modern Chinese Dictionary)*, 7 edition. The Commercial Press, Beijing.

Matteo Fuoli, Wen Huang, Jeannette Littlemore, Samuel Turner, and Emma Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning. *Preprint*, arXiv:2509.24866.

Hongyu Gong, Kshitij Gupta, Akriti Jain, and Suma Bhat. 2020. IlliniMet: Illinois system for metaphor detection with contextual and linguistic information. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 146–153. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Wei Huang and Ming Liu. 2026. Interpretable chinese metaphor identification via llm-assisted mipvu rule script generation: A comparative protocol study. *Preprint*, arXiv:2603.10784.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Lo"ic Barrault. 2023. A chinese metaphor corpus and its use for metaphor recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiaofei Lu and Ben Pin-yun Wang. 2017. Towards a metaphor-annotated corpus of Mandarin Chinese. *Language Resources and Evaluation*, 51(3):663–694.

Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3888–3898, Florence, Italy. Association for Computational Linguistics.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Susan Nacey, Lauge Greve, Aletta Dorst, and Tina Krennmayr, editors. 2019. *Metaphor Identification in Multiple Languages: MIPVU Around the World*. John Benjamins, Amsterdam.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Yujie Shao, Linquan Liu, Yanyan Lan, Lei Wang, and Tiejun Zhao. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds for metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins, Amsterdam.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39. Association for Computational Linguistics.

Ben Pin-yun Wang, Xiaofei Lu, Chih-Chung Hsu, Eric Po-Chung Lin, and Haiyang Ai. 2019. Metaphor identification in chinese. In Susan Nacey, Lauge Greve, Aletta Dorst, and Tina Krennmayr, editors, *Metaphor Identification in Multiple Languages: MIPVU Around the World*. John Benjamins, Amsterdam.

Shenglong Zhang, Ying Liu, and Yanjun Ma. 2021. SaGE: Syntax-aware GCN with ELECTRA for chinese metaphor detection. In *Proceedings of the 20th*

Chinese National Conference on Computational Linguistics, pages 667–677, Huhhot, China. Chinese Information Processing Society of China.

A MCD7 Construction Details

In the cross-reference resolution stage, 5,947 entries (7.95%) contain reference indicators (e.g., *jian* 'see', *tong* 'same as', *cankan* 'cf.') that redirect to other entries. We recursively resolve these references with a maximum depth of 5 and circular-reference detection. Of the 5,947 referencing entries, 4,861 (81.74%) resolve successfully; 1,081 fail due to missing target entries, and 5 form cycles.

B Qwen Task-Formulation Comparison

We previously reported (Table 3, single-seed) that Q2 generative JSON extraction is the strongest of six task formulations we tried for Qwen3.5-9B.

Two findings are robust across formulations:

1. **No Qwen formulation reaches encoder-baseline performance.** Even Q2 (0.6275) is approximately 0.087 below RoBERTa's mean (0.7142). The next-strongest formulation, Q1 token-level classification (0.5680 baseline; 0.5808 best after ablation), is about 0.13 below RoBERTa.
2. **Generation-style formulations dominate classification-style ones at the top of the ranking, but the two paradigms are not strictly ordered.** Q2 (generation) outperforms Q1 (classification) by 0.060, but Q1 in turn outperforms Q8 v2 (generation, longer context). The poorest two formulations are Q4 (BIO span, classification) and Q8 v1 (structured generation with truncated supervision).

Two failure modes are worth highlighting for their generalizability.

- **Q4 (BIO Span)** achieves 0.4049 overall F1 but F1 $= 0$ on the I-tag class. Chinese metaphors are predominantly single-token at the lexical-unit granularity used by PSU CMC; multi-token metaphor spans (which BIO tagging is designed to handle) are rare, and the I-tag head learns to predict near-zero probability everywhere.
- **Q8 v1 (max length 256)** collapses to F1 ≈ 0, while Q8 v2 (max length 512) recovers to 0.5299. The cause is supervision-token truncation: in Q8 v1, the model output is frequently cut off before the relevant tokens are generated.

C Qwen Precision–Recall Analysis

The Qwen Q2 generative configuration is roughly 11 absolute F1 points below the encoder baselines on PSU CMC test (Table 2). Crucially, this gap is concentrated in *recall* (0.5526 ± 0.0235) rather than precision (0.6963 ± 0.0103): Qwen tends to be conservative in flagging metaphor, missing borderline cases that the encoder baselines catch.

Two structural causes are likely. First, generative output forces a discrete commitment per token (the token either appears in the generated JSON list or it does not), with no calibrated probability gradient as in classification heads. Tokens whose metaphor status is genuinely ambiguous are systematically dropped rather than included with low confidence. Second, the QLoRA low-rank adaptation modifies only the four attention projection matrices and may underfit a fine-grained token-level signal compared to a dense classification head trained on top of full-rank hidden states. The combination—a hard discrete decision plus low-rank adaptation—biases Qwen toward the precision-favoring regime we observe.

D Per-Seed Results

Figure 3 shows per-seed test positive F1 for the four model configurations.

E Hyperparameter Configurations

Table 4 lists the final training hyperparameters for all model configurations.

Table 3: Test positive F1 of six Qwen3.5-9B + QLoRA task formulations on PSU CMC test set, single seed (=42).

Task Form	Type	Test pos-F1	Notes
Q2: Generative JSON	Generation	0.6275	best
Q1: Token CLS	Classification	0.5680	wd=0.01 best
Q8 v2: Structured Gen (max_len=512)	Generation	0.5299	after fix
Q6: QA-style	Generation	0.4915	short prompt
Q4: BIO Span	Classification	0.4049	I-label F1=0
Q8 v1: Structured (max_len=256)	Generation	0.0090	failed

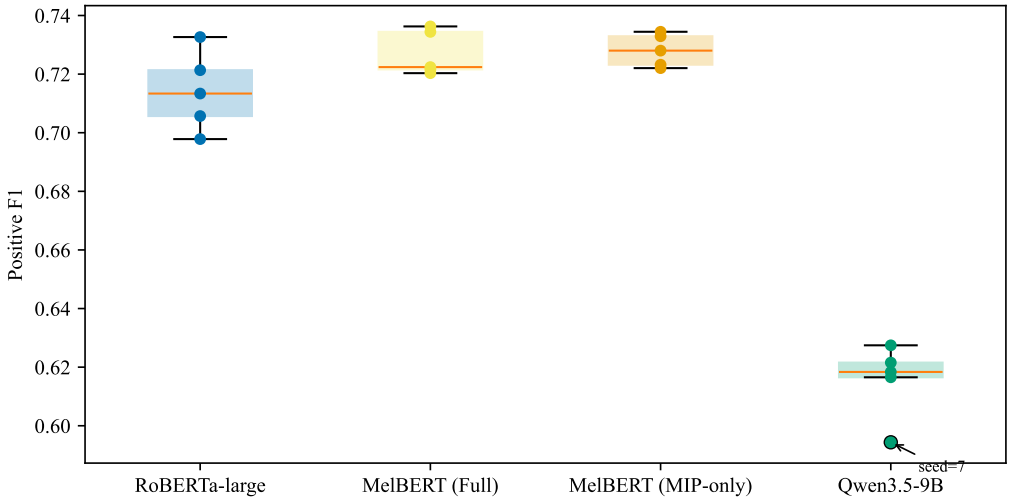

Figure 3: Per-seed test positive F1 for the four model configurations. Each point is one seed; box plots show median, IQR, and range across the 5 seeds. Qwen seed 7 is highlighted as a fiction-register outlier.

Table 4: Hyperparameters for the four model configurations.

Model	Backbone	LR	Batch (eff.)	Epochs	Max Length	Other
RoBERTa	hfl/chinese-roberta-wwm-ext-large	5e-5	16	10	256	early_stop_patience=3
MelBERT	hfl/chinese-roberta-wwm-ext-large	5e-5	16	10	256	full: MIP+SPV; early_stop_patience=3
Qwen Q2	Qwen3.5-9B	2e-4	1x16=16	3	1024	QLoRA qkvo, warmup=0.03
Qwen Q1	Qwen3.5-9B	1e-4	2x8=16	5	256	Token CLS, wd=0.01 best