Benchmarking Chinese Tenor–Vehicle Generation: Protocols, Baselines, and Semantic Evaluation

Yufeng Wu
City University of Hong Kong, Hong Kong
yufenwu2-c@my.cityu.edu.hk

Abstract

Generating a VT (Tenor–Vehicle) pair for a Chinese metaphor requires identifying its target and source concepts, which may extend beyond sentence substrings. We establish a dedicated benchmark for this task using 600 metaphor-positive records and existing annotations from CCIME. The benchmark uses a common output schema, failure-inclusive scoring, group-safe partitions, analysis by reference availability, and reference-blind human evaluation. Eight baseline conditions cover five prompting procedures, a span-constrained encoder, an unadapted generative base, and QLoRA (quantized low-rank adaptation). Direct prompting obtains a Pair Mean Char-F1 of 0.674; additional role guidance, decomposition, and candidate selection do not consistently improve recovery. Automatic scores are highest when both references occur on the sentence surface and lowest when neither does. In the prespecified matched comparison, QLoRA reaches 0.700 and improves over its base by +0.0165 (95% group-bootstrap interval [+0.0049, +0.0283]; Holm $p = 0.0223$). On 40 sampled sentences, both human assessors accept 34 QLoRA pairs, including 23 without exact reference agreement. The benchmark provides reference performance and diagnostic evidence for evaluating Chinese VT generation across lexical agreement, semantic acceptability, output constraints, and cost.

Keywords: Chinese metaphor, Tenor–Vehicle generation, benchmarking, structured generation, semantic evaluation

1 Introduction

Chinese VT (Tenor–Vehicle) generation makes a metaphor's target and source concepts explicit. Given a metaphorical sentence, a system returns the Tenor, the concept being described, and the Vehicle, the source concept used to characterize it. Either role may be directly expressed or left implicit. A benchmark must therefore evaluate ordered concept pairs while allowing descriptions beyond the sentence's exact wording.

This task combines role assignment with concept formulation. A system must identify the concept being characterized and the concept supplying the comparison, then express each role. Recovering the right words in the wrong order changes the interpretation. Likewise, a concept can be appropriate even when its wording differs from the annotation. These properties make sentence-to-pair generation a useful setting for examining the relationship between textual recovery and semantic interpretation.

Existing resources provide task foundations. CMRE models metaphorical span relations (Chen et al., 2023); CERD evaluates rhetorical components and generation (Liu et al., 2024); and CMDAG supplies Tenor, Vehicle, and Ground annotations (Shao et al., 2024). Generative systems already return rhetorical components through prompted or adapted models (Lai et al., 2025). CCIME includes Tenor and Vehicle recovery within metaphor identification (CCIME 2026 Organizers, 2026). These precedents motivate a dedicated evaluation of sentence-to-pair generation.

To our knowledge, this is the first dedicated benchmark study of Chinese free-text Tenor–Vehicle generation from metaphorical sentences. We build it using 600 metaphor-positive CCIME records and their existing annotations. It specifies the free-text pair target, record identifiers, grouped partitions, output schema, lexical metrics, and treatment of failed outputs. Four categories defined by reference availability expose output-space constraints, and human evaluation assesses interpretations beyond exact reference agreement.

We ask how representative prompting, extraction, and adaptation methods perform on Chinese VT generation and how reference availability

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 171–179
November 6-7, 2026 ©2026 Association for Computational Linguistics

affects their evaluation. Eight conditions cover five prompting procedures, a span-constrained encoder, an unadapted generative base, and QLoRA (quantized low-rank adaptation). Prespecified paired comparisons test differences between selected methods; cross-family scores describe systems with different training access and output spaces.

The study contributes a dedicated evaluation protocol, results for eight baseline conditions, and an analysis of reference agreement alongside independent human judgments. The comparisons characterize the evaluated prompting procedures and task adaptation, while the human judgments identify acceptable outputs that differ from the official reference.

2 Related Work

Related work addresses metaphor interpretation, rhetorical-component extraction, and generation from specified concepts. These tasks differ in their inputs and outputs, which determines their relevance to sentence-to-pair generation.

2.1 Structured Metaphor Recovery

Metaphor identification and interpretation address different outputs (Shutova, 2010). Contextual and syntax-aware identification models (Gao et al., 2018; Mao et al., 2019; Choi et al., 2021; Zhang et al., 2021) locate figurative uses. Explanation resources make interpretations explicit. Li et al. (2013) separate recognition from explanation, Figure Me Out provides reference interpretations (Zayed et al., 2020), and MUNCH evaluates several dimensions of understanding (Tong et al., 2024).

Chinese component resources offer closer task foundations. CMRE predicts typed span relations (Chen et al., 2023). CERD defines component extraction over sentence-token indices and compares encoders, prompted LLMs (large language models), and models adapted with LoRA (low-rank adaptation) (Liu et al., 2024). Lai et al. (2025) generate JSON rhetorical components, including Tenor and Vehicle, using LoRA and in-context learning. NLPCC component tasks also support option-based evaluation (Wang et al., 2024). Our benchmark evaluates one ordered free-text pair per sentence, including reference forms absent from the input, using the same pair-level metrics for all baselines.

Generation from specified components provides a complementary direction. CMDAG studies metaphor generation from Tenor–Vehicle or Tenor–Ground inputs (Shao et al., 2024). Our input is an existing metaphorical sentence and the output is its concept pair. Analogy extraction likewise permits explicit and inferred elements (Boisson et al., 2025). These tasks motivate evaluating both reference agreement and the semantic acceptability of inferred concepts.

2.2 Prompting and Structured Generation

Prompting provides several inference-time routes to structured recovery. Intermediate reasoning can expose a sequence of decisions (Wei et al., 2022); decomposition makes those decisions explicit; and repeated generation followed by selection introduces search over candidate outputs. Metaphor evaluation can also vary with prompt configuration and surface properties (Sanchez-Bayona and Agerri, 2025). We compare compact instructions, role definitions, decomposition, full task specifications, and candidate selection to examine differences among these inference procedures.

Structured generation also separates interface validity from content recovery. Grammar-constrained decoding can improve syntactic validity (Geng et al., 2023), but valid JSON can still contain reversed roles, unsuitable boundaries, or the wrong concepts. Reporting content quality alongside schema validity, while retaining invalid outputs in the denominator, captures both properties.

2.3 Task Adaptation and Output-Space Constraints

Task adaptation changes model parameters rather than only the inference instruction. QLoRA combines four-bit quantization with trainable low-rank adapters while keeping the base checkpoint frozen (Dettmers et al., 2023). In VT recovery, such adaptation can expose a model to the ordered schema and official reference conventions. A matched base condition separates adaptation gains from differences in the checkpoint or prompt.

Output space creates a separate representational constraint. Span-based systems are aligned with textual evidence, whereas free-string generators can return canonical or implicit concepts. Lexical overlap and sentence length can affect apparent metaphor understanding (Sanchez-Bayona and Agerri, 2025), motivating analysis by reference availability. We therefore compare span extraction with free-text generation and assess their outputs

using both reference matching and human judgments.

3 VT Generation Benchmark

The benchmark maps a metaphor-positive Chinese sentence x to one ordered pair (t, v) of free-text role descriptions. Either string may extend beyond substrings of x. CCIME supplies the annotations. We specify the evaluated subset, partitions, output schema, and scoring procedure below.

3.1 Benchmark and Ordered Target

The benchmark uses 600 metaphor-positive Chinese sentences from CCIME, retaining one official Tenor and one official Vehicle per record (CCIME 2026 Organizers, 2026). Record identifiers were fixed before inference and checked with a SHA-256 digest. Connected components of normalized exact-match links and near-duplicate links under fixed character-trigram and edit-similarity criteria define 599 resampling groups.

The task assumes a metaphorical input sentence. Each prediction describes one annotated relation through its ordered roles; the evaluation unit is the complete sentence–pair record. This separates the benchmark from a pipeline that first identifies metaphorical words or determines whether a sentence contains a metaphor. In concept generation, the system constructs both role strings instead of selecting a supplied option. The scoring procedure checks each string in its assigned role.

The existing outer assignment contains five folds of 120 records. Every connected group occurs in only one outer fold. Prompt methods do not train on the benchmark and therefore run once over all records. The structured encoder and QLoRA produce one OOF (out-of-fold) prediction per record. OOF denotes predictions made only for records excluded from model training. Text-free reproducibility manifests record identifiers, groups, fold assignments, and manifest checksums without redistributing the source sentences.

3.2 Shared Evaluation Contract

All eight conditions are converted to one canonical record schema with identifier, fold, Tenor, Vehicle, parser status, model identity, prompt hash, and provenance. At test time, only the sentence enters the model request; its official roles are withheld. Supervised methods use role annotations from their training partition. A malformed or terminal output remains in the 600-record denominator and receives zero content score.

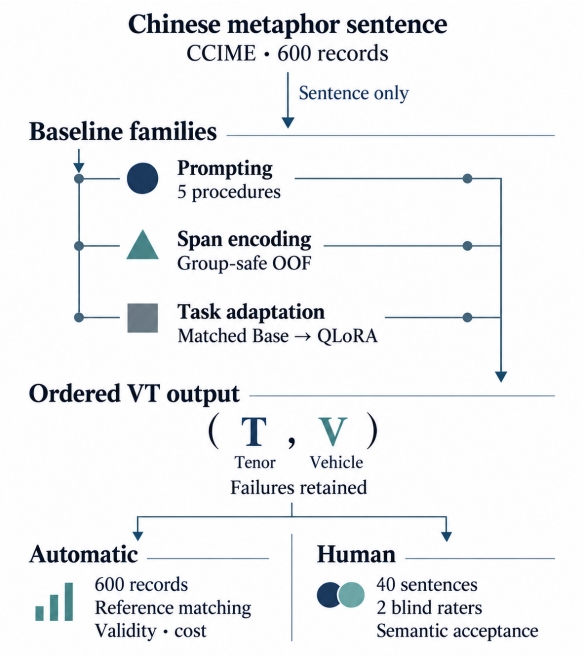

Figure 1: Chinese Tenor–Vehicle generation benchmark: task, baseline families, and complementary automatic and human evaluation.

The common record schema provides a scoring interface for methods with different native outputs. An encoder supplies extracted spans, while a generator supplies role strings through its response parser. Both are evaluated as ordered text fields after conversion. This allows the benchmark to report content recovery and interface reliability together. In particular, validity records output-schema compliance, while the role scores describe how well the resulting fields agree with the reference.

Figure 1 traces the pipeline from sentence input to a scored ordered VT record. Prompt procedures and Matched Base use full-coverage inference without benchmark training, while the structured encoder and QLoRA produce group-safe OOF predictions. Outputs from all three method families are converted to the same record schema and scored by the same evaluator. For each evaluated sentence, its official reference enters only at scoring.

3.3 Reference Baselines

We compare inference-time prompting, span-constrained encoding, and generative recovery with and without task adaptation. All conditions are scored with the same evaluator.

The three families examine different choices in VT recovery. Prompting changes the instructions

and inference procedure; the encoder tests recovery within a sentence-span output space; and the matched generative pair tests learning from task annotations. The five prompt conditions vary role guidance, intermediate decomposition, task specification, and candidate selection. These conditions provide baseline results for the specified inference procedures. The Matched Base–QLoRA pair holds the generative checkpoint and task interface constant while changing task-specific parameter adaptation.

3.3.1 Inference-Time Prompting

Direct uses a compact instruction that names the two roles and requires a strict two-field JSON object. Role-Guided defines the target and source roles in a fixed interpretation instruction and returns Tenor, Vehicle, and an interpretation field; the last is retained but not scored. Its original parser is preserved.

Decomposed requests candidate roles, role-order and boundary checks, and a final pair in one structured response. Full Skill adds guidance on surface preference, implicit concepts, minimal sufficient boundaries, and selection of one central mapping. Direct, Role-Guided, Decomposed, and Full Skill use the same quantized `qwen3.6:27b-q4_K_M` model tag and digest, Ollama endpoint, and deterministic decoding configuration. Direct, Decomposed, and Full Skill share the sentence-only payload and two-field parser, whereas comparisons involving Role-Guided evaluate the complete procedure, including its original output schema and parser. The underlying model is the official `Qwen/Qwen3.6-27B` release (Qwen Team, 2026).

Multi-Candidate runs the Full Skill generator five times with fixed stochastic seeds and temperature 0.7. Valid unique pairs are anonymized and deterministically shuffled before a blinded selector chooses one candidate verbatim. Candidate labels and order contain no seed or quality signal, and the selector cannot merge or rewrite candidates. A fixed fallback selects the first anonymized valid candidate if selection fails.

3.3.2 Span-Constrained Structured Encoding

The structured encoder uses the Chinese RoBERTa backbone `chinese-roberta-wwm-ext-large` with Tenor and Vehicle start and end heads plus two surface-presence heads. It emits a sentence span or null for each role, and non-surface targets mask their span loss. Checkpoint selection uses inner-validation Pair Mean Char-F1, then joint exact recovery, then the earlier epoch. The five-fold predictions cover every record once under the group-safe OOF design.

3.3.3 Generative Base and Task Adaptation

Matched Base uses the unadapted `Qwen/Qwen3.6-27B` checkpoint (Qwen Team, 2026), the fixed two-field training prompt, deterministic inference over all 600 sentences, and no repair or format retry. QLoRA uses the same checkpoint and prompt with NormalFloat four-bit quantization, double quantization, bfloat16 computation, and low-rank adapters (Dettmers et al., 2023). Adapter rank is 8, scaling is 16, dropout is zero, and bias adaptation is disabled.

Training uses AdamW at a constant learning rate of 10^{-4}, micro-batches of one with 16-step gradient accumulation, a maximum sequence length of 256, and seed 20260802. Loss applies only to the target JSON; adapters target 304 language-attention projections.

Within each outer fold, 432 records train candidate adapters and 48 group-safe records select among at most three epochs. Selection maximizes Pair Mean Char-F1, then joint exact recovery, then the earlier epoch. The selected epoch count is refit on all 480 outer-training records before inference on the untouched 120-record outer fold. Four folds select two epochs and one selects one epoch. Only LoRA adapter parameters are trainable, while the base-model parameters remain frozen.

3.4 Automatic Evaluation

Every output is converted to the ordered record schema before scoring. The evaluator applies Unicode NFKC (Normalization Form Compatibility Composition) normalization and removes whitespace. It reports exact match, multiset character-overlap F1, and bidirectional substring containment for each role. Pair Mean Char-F1 is the arithmetic mean of the Tenor and Vehicle character F1 values for each record and then across records. These measures quantify lexical agreement with one reference. Human evaluation separately assesses semantic acceptability. Malformed and terminal outputs remain in the 600-record denominator and receive zero content score.

The metrics capture different degrees of agreement. Exact match requires the normalized role strings to coincide with their references. Character F1 gives partial credit for shared characters, including overlap between differently bounded descrip-

tions. Joint exact and joint containment require both ordered roles to satisfy their field-level criterion. Success on one field cannot compensate for failure on the other. Pair Mean Char-F1 preserves graded information from both fields. Reporting these measures together distinguishes partial lexical recovery from complete ordered-pair recovery.

Five contrasts were prespecified: Role-Guided versus Direct, Decomposed versus Role-Guided, Full Skill versus Role-Guided, Multi-Candidate versus Full Skill, and QLoRA versus Matched Base. Confidence intervals use 10,000 paired group-bootstrap resamples stratified by outer fold. Two-sided group sign-flip tests use 100,000 draws over 599 connected groups, with Holm correction over all five contrasts. The analysis uses seed 20260802 and deterministic byte serialization for exact reproduction. The supporting document provides the verbatim prompts, the Role-Guided archive-and-completion provenance, prompt-compression records, full cost ledger, mechanical outcome taxonomy, QLoRA details, fold manifests, and reproducibility hashes.

3.5 Reference-Blind Human Semantic Audit

Two master's students who were native speakers of Chinese independently judged Direct, Encoder, Matched Base, and QLoRA on the same 40 sentences. Ten sentences were sampled from each reference-surface category, with 40 distinct groups. Identical ordered outputs within a sentence were deduplicated, yielding 103 items per assessor. Methods, official references, and automatic scores were hidden.

Equal sampling across the four categories gives the audit direct coverage of both surface-supported and non-surface reference conditions. Evaluating the same sentences for each method keeps sentence content fixed across the comparison. Deduplication presents an identical pair only once within a sentence, so repeated outputs do not create repeated rating decisions. The resulting judgments are then associated with the methods that produced each pair. Blinding directs attention to the sentence and proposed interpretation rather than model identity or lexical agreement with the reference.

Assessors rated each role and the ordered pair as acceptable (1), unacceptable (0), or uncertain (U); pair acceptance required both roles and their correspondence to be acceptable. All outputs remained in the denominator. We retained U without adjudication.

After receiving the ratings, we additionally reported population-weighted acceptability bounds: the lower bound counts accepted pairs, and the upper also counts U, weighting categories by their proportions among the 600 records. For category h, let N_h be its benchmark size and a_h, u_h the accepted and uncertain counts among its ten sampled sentences. The bounds are $L = \sum_h (N_h/600)(a_h/10)$ and $H = \sum_h (N_h/600)((a_h + u_h)/10)$, calculated separately for each method and assessor. They express judgment uncertainty, not sampling confidence. The audit is descriptive.

4 Benchmark Results

We report overall performance and operational cost, test prespecified method contrasts, examine variation with reference availability, and assess semantic acceptability through independent human judgments.

4.1 Overall Results

Table 1 summarizes recovery quality and output validity across all eight conditions. Direct has the highest observed prompting Pair Mean Char-F1 (0.674), followed by Multi-Candidate (0.670); additional prompt structure does not consistently improve recovery. The generative comparators score 0.684 for Matched Base and 0.700 for QLoRA, while the span-constrained encoder scores 0.464. These cross-family scores describe performance under the reported training conditions and output constraints.

Output validity and reference recovery show different patterns. Role-Guided, Full Skill, and Multi-Candidate produce fully valid outputs, but none exceeds Direct's observed recovery score. Direct and Decomposed retain scored failures. Thus, complete schema compliance does not identify the strongest reference recovery in this benchmark.

The role-level scores add detail to the pair averages. In Table 1, every condition has higher Vehicle character F1 than Tenor character F1. This asymmetry is especially pronounced for the span encoder. The separate role F1 scores identify an imbalance that a single pair average would obscure. Joint containment provides a complementary measure of whether both ordered fields satisfy the substring criterion.

Operational cost also differs. Direct requires 615 calls and 453.1 seconds, Full Skill 600 calls

and 467.4 seconds, and Multi-Candidate 3603 calls and 2710.1 seconds. These Ollama totals include retries and sum call durations; candidate search increases cost without a supported gain over Full Skill. Matched Base and QLoRA take 1270.2 and 1608.5 seconds in Transformers, including model or adapter loading but excluding training. Timing comparisons apply within each runtime because the timing scopes differ.

Panel A: paired recovery and validity

Method	Pair F1 [95% CI]	Exact	Valid
Direct	0.674[0.651,0.698]	0.275	0.975
Role-Guided	0.662[0.640,0.684]	0.230	1.000
Decomposed	0.635[0.611,0.658]	0.213	0.980
Full Skill	0.662[0.640,0.684]	0.213	1.000
Multi-Candidate	0.670[0.649,0.691]	0.225	1.000
Encoder	0.464[0.437,0.491]	0.163	1.000
Matched Base	0.684[0.661,0.706]	0.278	0.995
QLoRA	0.700[0.678,0.723]	0.298	1.000

Panel B: role scores and pair containment

Method	Tenor F1	Vehicle F1	Containment
Direct	0.619	0.730	0.547
Role-Guided	0.613	0.711	0.493
Decomposed	0.576	0.693	0.523
Full Skill	0.616	0.709	0.503
Multi-Candidate	0.623	0.718	0.498
Encoder	0.309	0.619	0.320
Matched Base	0.622	0.746	0.535
QLoRA	0.644	0.757	0.547

Table 1: Reference performance on 600 records. F1 denotes character F1; Exact and Containment require both roles. Valid is schema validity. Intervals are group-bootstrap intervals. Encoder and QLoRA use OOF predictions; other conditions use full-coverage inference.

4.2 Prompting and Task Adaptation

Figure 2 distinguishes observed score differences from the five prespecified paired tests. Two contrasts remain significant after Holm correction. Decomposed minus Role-Guided is -0.0275 in Pair Mean Char-F1 (95% interval [-0.0406, -0.0146], $p = 0.0001$). QLoRA improves over Matched Base by +0.0165 (95% interval [+0.0049, +0.0283], $p = 0.0223$), the only positive contrast significant after correction.

The intervals for Role-Guided versus Direct, Full Skill versus Role-Guided, and Multi-Candidate versus Full Skill include zero. The paired evidence therefore supports a gain from task adaptation in this configuration and lower recovery for Decomposed than Role-Guided, while the other three prompting comparisons remain inconclusive.

Figure 2 also shows why the observed ranking and the paired tests answer different questions. A higher mean supplies a descriptive ordering, whereas a contrast examines the difference on matched records under the specified resampling procedure. For the three intervals crossing zero, the estimated difference remains compatible with either method scoring higher.

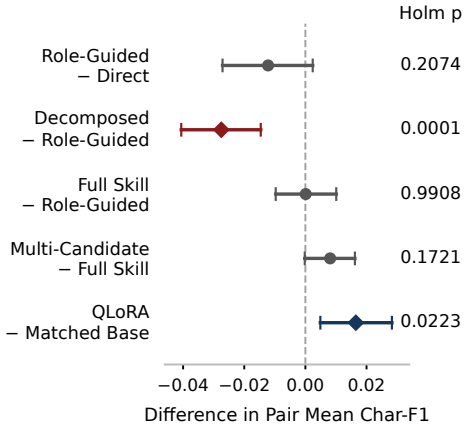

Figure 2: Five prespecified paired differences in Pair Mean Char-F1. Points show differences (first method minus second); lines show 95% group-stratified bootstrap intervals. The dashed line marks zero. Holm-corrected group sign-flip p values appear at right; diamonds mark $p < 0.05$.

4.3 Performance by Reference Availability

Table 2 groups sentences by whether each normalized official role string occurs as an input substring: both roles (203), Tenor only (51), Vehicle only (170), or neither (176). Reference availability denotes this observable string relation. Every condition scores highest when both references are present and lowest when neither is present. QLoRA ranges from 0.950 to 0.415 across these two categories; the encoder ranges from 0.800 to 0.145.

Free-text generators retain more reference overlap than the encoder in the neither-surface category. This descriptive gradient relates recovery scores to the availability of the official strings in the input.

Vehicle-only scores exceed Tenor-only scores for seven conditions. Full Skill reverses this pattern, with 0.673 for Tenor only and 0.655 for Vehicle only. The two categories therefore reveal variation that would be lost by combining all sentences with one surface reference. The neither-surface category is particularly informative about output constraints. A free-text method can formulate a role beyond the input substrings, whereas the encoder remains restricted to those substrings or null outputs.

Method	Both (203)	T only (51)	V only (170)	Neither (176)
Direct	0.944	0.594	0.681	0.381
Role-Guided	0.914	0.617	0.659	0.389
Decomposed	0.902	0.615	0.639	0.328
Full Skill	0.899	0.673	0.655	0.393
Multi-Candidate	0.910	0.646	0.662	0.409
Encoder	0.800	0.249	0.456	0.145
Matched Base	0.941	0.650	0.690	0.391
QLoRA	0.950	0.645	0.715	0.415

Table 2: Pair Mean Char-F1 by reference-surface category. T and V indicate official Tenor and Vehicle strings occurring as normalized input substrings. Counts appear once in the headers.

4.4 Human Evaluation

Table 3 evaluates whether the generated pairs are acceptable interpretations of the same 40 sentences. Both assessors accept 34 QLoRA pairs, 29 Matched Base pairs, 28 Direct pairs, and 10 Encoder pairs. Across the 103 unique rated outputs, pair judgments agree on 102 (99.0%; three-category Cohen's $\kappa = 0.982$). All 98 outputs receiving binary judgments from both assessors agree; five receive U from at least one assessor.

Semantic acceptance extends beyond exact reference recovery. 23 of QLoRA's 34 accepted pairs are nonexact, as are 20 of 29 for Matched Base, 18 of 28 for Direct, and 5 of 10 for Encoder. Thus, nonexact outputs include acceptable alternatives under the assessors' judgments. These paired sample results are descriptive; per-role agreement and category-level counts are reported in the supporting document.

The weighted bounds in Table 3 answer a different question from the raw accepted counts. The counts summarize the equally stratified audit sample, while the bounds weight each category according to its frequency in the benchmark. Their lower endpoints include accepted judgments, and their upper endpoints additionally include uncertain judgments. The separate assessor bounds preserve differences in uncertainty even when accepted counts agree. Unresolved pairs retain their uncertain status in the reported bounds.

Method	Accepted A=B	U A/B	Weighted bounds (%) A/B
Direct	28	2/2	A=B: 75.52–81.38
Encoder	10	0/0	A=B: 26.23–26.23
Matched Base	29	2/2	A=B: 76.37–82.23
QLoRA	34	1/2	A: 88.65–91.58 B: 88.65–94.52

Table 3: Human pair judgments on 40 sentences per method. A and B denote the two assessors. Accepted counts agree across assessors; U counts and population-weighted acceptability bounds are shown separately. Bounds retain U and exclude sampling uncertainty.

5 Discussion

The results distinguish the performance of the prompting procedures, the constraints of span extraction, and the contribution of task adaptation. Human judgments clarify how reference scores relate to acceptable interpretations.

5.1 What the Benchmark Reveals

Direct has the highest observed prompt score, while elaboration, decomposition, and candidate search do not consistently improve recovery. The benefit of intermediate reasoning (Wei et al., 2022) therefore requires task-specific evaluation. Output validity is a separate property. Grammar constraints protect the interface (Geng et al., 2023), but role selection and conceptual correspondence remain semantic decisions after valid JSON has been produced.

The reference-surface gradient exposes an output-space constraint. A span encoder cannot return an official reference string absent from its input sentence. Free-text generators can express such concepts, although their scores also depend on the chosen reference formulation. This distinction connects component extraction (Chen et al., 2023; Liu et al., 2024) to concept-pair generation. The categories therefore describe the relationship between the reference strings and each method's output space.

The accepted nonexact pairs in the human audit show why semantic assessment complements reference matching. Pair acceptance requires both suitable role descriptions and a context-supported correspondence between them. Five unique outputs received an uncertain pair judgment from at least one assessor despite high agreement. Retaining these judgments distinguishes unresolved interpretations from errors.

The intended output guides system choice. Span

encoders suit textual-component targets, while prompting and adapted generators can return free-text concepts.

Benchmark users can interpret results through complementary evaluations. Overall recovery and validity establish whether a method returns usable ordered predictions. Reference-surface results indicate which input–reference conditions contribute to its performance. Human judgments then distinguish unsuitable interpretations, acceptable alternatives, and unresolved cases among outputs that disagree with the official strings. Cost describes the resources needed to obtain those predictions. Keeping these dimensions separate clarifies whether an improvement concerns lexical agreement, semantic acceptance, response reliability, or inference effort.

5.2 A Reference Point for Task Adaptation

The Matched Base–QLoRA comparison isolates task adaptation more closely than the cross-family score rankings by sharing the checkpoint, prompt, output schema, and deterministic inference. Its gain supplies a parameter-efficient adaptation baseline (Dettmers et al., 2023) for ordered VT generation under the reported protocol.

The comparison is consistent with learning the ordered output and reference conventions from annotations. The human audit complements this result with semantic acceptability judgments. Together, these observations support evaluating adaptations on both reference agreement and meaningful concept correspondence.

The matched comparison provides a starting point for subsequent benchmark studies. A new adaptation can be compared with the unadapted checkpoint using the same inputs, output fields, parser, and grouped evaluation. Changes in any of these components should be reported alongside the scores because they specify training access, permitted outputs, and how predictions are scored. This use of the benchmark preserves the distinction between a task-adaptation effect and an improvement obtained by changing the evaluation interface.

6 Limitations

The study covers one 600-item Chinese benchmark and one generative model family, so its comparative claims apply to the reported conditions rather than unseen datasets or a universal method ordering. Each record has one official ordered pair; exact match and character overlap can penalize plausible alternatives or reward surface similarity without establishing conceptual equivalence. Reference-surface categories describe string occurrence. Syntactic types and conceptual implicitness would require separate annotations.

Prompt procedures, the span encoder, and QLoRA also differ in training access and output constraints, which limits cross-family values to descriptive comparison. One prespecified training seed does not estimate adaptation variance. The group-bootstrap interval describes benchmark-group variation conditional on the fitted fold models. Assessing training variability requires independently repeated training seeds.

The human audit covers only 40 sentences and four methods. Uncertain judgments were not adjudicated, and the reported bounds quantify judgment uncertainty rather than sampling variability.

7 Conclusion

We established a dedicated benchmark for Chinese VT generation using 600 CCIME records, existing role annotations, grouped evaluation, and a shared failure-inclusive scoring procedure. Eight baseline conditions provide reference performance across prompting, span-constrained encoding, and task adaptation. Prompt elaboration does not consistently improve recovery, reference-surface categories expose output-space constraints, and matched QLoRA improves over its unadapted base by +0.0165. The human audit identifies accepted nonexact pairs while preserving uncertain interpretations. These results make the benchmark useful for assessing generated concept pairs through reference agreement, semantic acceptability, output reliability, and operational cost.

References

Joanne Boisson, Zara Siddique, Hsuvas Borkakoty, Dimosthenis Antypas, Luis Espinosa Anke, and Jose Camacho-Collados. 2025. Automatic extraction of metaphoric analogies from literary texts: Task formulation, dataset construction, and evaluation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6692–6704. Association for Computational Linguistics.

CCIME 2026 Organizers. 2026. CCL2026-CCIME Chinese conversational implicature and metaphor capability evaluation. Official evaluation repository. Accessed 2026-07-31.

Guihua Chen, Tiantian Wu, MiaoMiao Cheng, Xu Han, Jiefu Gong, Shijin Wang, and Wei Song. 2023. Chinese metaphorical relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9085–9095. Association for Computational Linguistics.

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 NAACL-HLT*, pages 1763–1773.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952.

Yuxuan Lai, Xiajing Wang, and Chen Zheng. 2025. System report for CCL25-eval task 6: Chinese essay rhetoric recognition using LoRA, in-context learning and model ensemble. In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025)*, pages 220–232. Chinese Information Processing Society of China.

Hongsong Li, Kenny Q. Zhu, and Haixun Wang. 2013. Data-driven metaphor recognition and explanation. *Transactions of the Association for Computational Linguistics*, 1:379–390.

Nuowei Liu, Xinhao Chen, Hongyi Wu, Changzhi Sun, Man Lan, Yuanbin Wu, Xiaopeng Bai, Shaoguang Mao, and Yan Xia. 2024. CERD: A comprehensive Chinese rhetoric dataset for rhetorical understanding and generation in essays. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6744–6759. Association for Computational Linguistics.

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*, pages 3888–3898.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model. Official model release documentation. Model identifier Qwen/Qwen3.6-27B; accessed 2026-08-03.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17462–17477. Association for Computational Linguistics.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Stephen W. Huang, Ge Zhang, and Jie Fu. 2024. CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 3357–3366.

Ekaterina Shutova. 2010. Models of metaphor in NLP. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 3517–3536.

Jie Wang, Jin Wang, and Xuejie Zhang. 2024. Chinese metaphor recognition using a multi-stage prompting large language model.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure Me Out: A gold standard dataset for metaphor interpretation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5810–5819.

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.