KoMMERCE: A Korean Multimodal and Multitask Evaluation Benchmark for Reliable Conversational E-commerce Agents

Byunggill Joe[1], San Kim[1], Minyoung Jung[1]
DONG HYOK SHIN[1], Yu-Jeong Sung[2]*, Saim Shin[1]
[1]Korea Electronics Technology Institute
[2]Kyung Hee University
byunggill@keti.re.kr, kimsan0622@keti.re.kr
minyoung.jung@keti.re.kr, sdh876@keti.re.kr
wjddb1025@khu.ac.kr, sishin@keti.re.kr

Abstract

E-commerce conversations are increasingly multimodal, yet end-to-end response scores cannot localize which dialogue competence failed, and no public benchmark evaluates fine-grained multimodal dialogue understanding for Korean commercial agents. We present KoMMERCE, a Korean multimodal fashion-domain benchmark that coordinates nine fine-grained dialogue understanding tasks on the same conversations: sentiment, intent, and strategy classification; image and text knowledge retrieval; dialogue state tracking; span detection; coreference resolution; and the first formulation of cross-modal dialogue relation extraction. The corpus contains 15,293 dyadic role-play dialogues under two-stage quality control. Baselines on six frontier multimodal LLMs reveal task-conditional patterns that should inform deployment choices: no model wins all nine tasks, and context length affects accuracy differently from one task to another.

1 Introduction

Real-world e-commerce conversations are increasingly multimodal: users send product photos, agents reply with image recommendations, and visual attributes routinely cross between text and image. LLM-powered commercial agents are rapidly entering these conversations through consumer-scale search, recommendation, and even payment execution (Gartner, Inc., 2025), placing high-stakes transactions under their control. Reliability evaluation must therefore move beyond text-only interactions and a single end-to-end response score, toward per-competence multimodal audit that localizes which dialogue understanding capability actually failed.

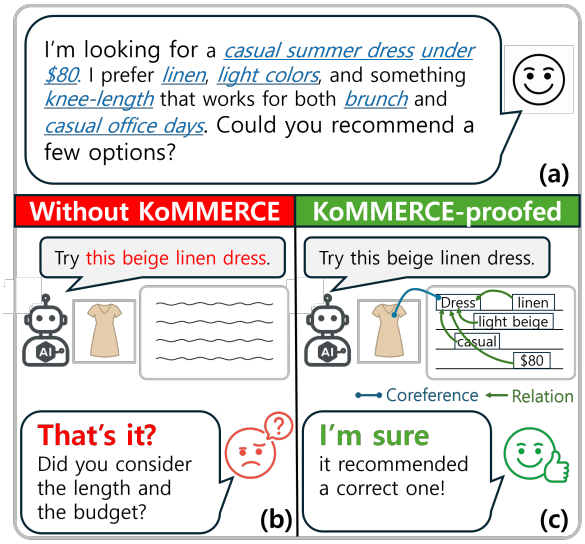


Figure 1: (a) User request with constraints. (b) A bare recommendation that hides which competence the agent used. (c) With KoMMERCE's nine NLU annotations (illustrated: span, coreference, and relation evidence), the recommendation becomes interpretable across each dialogue understanding step.

Consider a shopper asking for a beige linen dress (Figure 1a): a commercial agent's bare recommendation (Figure 1b) tells us neither what it understood about "beige linen" nor whether its visual grounding matches the textual constraints. This opacity is a barrier to reliable shopping automation: when the agent fails, neither the user nor the developer can tell which competence broke, so failures cannot be diagnosed or corrected at scale. Auditing each step (the spans the agent parsed, entities linked, visual attributes grounded against) turns an opaque recommendation into an interpretable one (Figure 1c).

*Work done while Yu-Jeong Sung was an intern at Korea Electronics Technology Institute.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4118–4150
November 6-10, 2026 ©2026 Association for Computational Linguistics

We propose KoMMERCE[1], a Korean multimodal benchmark over the fashion domain that addresses an acute gap: no public benchmark evaluates fine-grained multimodal dialogue understanding for Korean commercial agents (Section 2), even though Korea's e-commerce market (Statista Market Insights, 2025) subjects domestic LLM-based agents to the same reliability stakes outlined above. KoMMERCE coordinates nine fine-grained NLU tasks on the same dialogues: sentiment, intent, and strategy classification; image and text knowledge retrieval; dialogue state tracking; span detection; coreference resolution; and, for the first time to our knowledge, cross-modal dialogue relation extraction. Fashion concentrates the properties that stress multimodal dialogue understanding (dense visual attributes, frequent cross-turn references like "this in beige" or "a shorter version", and representative consumer e-commerce traffic) in a single domain. KoMMERCE dialogues are collected via dyadic human role-play grounded in K-Fashion images (AI-Hub, 2020) and curated Korean text knowledge, with every sample passing a two-level quality review that mitigates artifacts commonly attributed to crowd-sourced collection (details in Section 4).

We benchmark six frontier multimodal LLMs on KoMMERCE. Beyond what an end-to-end response score would surface, our fine-grained multimodal evaluation reveals two task-conditional failure patterns: longer context helps DST and retrieval but collapses coreference and relation extraction, and image context helps shallow classification while disrupting structured state tracking.

Our contributions are:

- KoMMERCE, a Korean multimodal benchmark that holistically evaluates LLM-agent dialogue understanding for e-commerce by coordinating *nine* fine-grained NLU tasks on the same conversations.
- Multimodal dialogue relation extraction, introduced here as a new task capturing entity–attribute links jointly grounded in images and conversational context.
- Baselines on six frontier multimodal LLMs that diagnose three failure patterns of current LLM-agents: no model wins all nine tasks, context length is task-conditional, and structured-task performance surfaces as a concrete improvement direction for the open-weight reference (Qwen3-VL-32B).

[1]Dataset and annotation documentation: https://github.com/MMC-K/multimodal_fashion_dialog_dataset

2 Related Work

KoMMERCE sits at the intersection of three orthogonal axes of prior work (multimodal dialogue benchmarks, fine-grained NLU benchmarks, and Korean LLM benchmarks), and further introduces cross-modal dialogue relation extraction, a task formulation not addressed by any prior resource we surveyed. A condensed comparison of representative resources along these axes is provided in Appendix F (Table 27).

Multimodal dialogue benchmarks. Existing multimodal dialogue corpora include early fashion-focused work (Saha et al., 2018), SIMMC 2.0 (Kottur et al., 2021) with state tracking, coreference, and disambiguation, MMDialog (Feng et al., 2023), and recent conversational recommendation in apparel (Wang et al., 2025b; Liao et al., 2021; Pang et al., 2025). These resources are English-dominant and seldom co-evaluate *coordinated* fine-grained NLU subtasks on the same dialogue or enforce explicit cross-modal relation annotations.

Fine-grained NLU benchmarks. Fine-grained NLU has been rigorously benchmarked outside the dialogue setting: text-level entity and coreference (Tjong Kim Sang and De Meulder, 2003; Weischedel et al., 2013; Pradhan et al., 2013), relation extraction (Zhang et al., 2017; Yao et al., 2019), vision–language grounding (Lin et al., 2014; Plummer et al., 2017; Wu et al., 2021), and cross-modal extraction (Zheng et al., 2021; He et al., 2023; Huang et al., 2025). These benchmarks evaluate each capability in isolation and omit cross-modal reasoning *across dialogue turns*; they also rarely support fine-grained fashion retrieval (pattern, length, sleeve style, neckline), where KoMMERCE annotates 686 detailed attributes.

Korean LLM benchmarks. Korean LLM benchmarks each cover a slice disjoint from our setting: KMMLU (Son et al., 2024) and KMMLU-Redux/Pro (Hong et al., 2025) for expert exam knowledge, KoCommonGEN v2 (Seo et al., 2024) for commonsense, and KoBALT (Shin et al., 2025) for linguistic phenomena. All three are single-turn, text-only multiple-choice and do not evaluate dialogue, multimodal grounding, or co-evaluated NLU. ECom-Bench (Wang et al., 2025a) is the

Task	Label	Description
Intent classification	REQUEST	The user asks for an initial recommendation of a specific item.
	REFINED-REQUEST	A follow-up request that narrows the item's attributes after the system's suggestion.
	EVALUATION	The user expresses an opinion about the system's suggestion.
	DETAIL	A request for detailed information about the requested fashion item.
	KNOWLEDGE	Utterances that ask for general, fashion-related knowledge.
	CHAT	Greetings, thanks, or small talk that does not require fashion items.
Strategy classification	CLARIFICATION	Questions that clarify the user's requirements.
	SUGGESTION	System suggestions tailored to the user's stated needs.
	DETAIL	Responses that provide specific details of the user-requested item.
	KNOWLEDGE	Fashion-related knowledge that answers the user's question.
	CHAT	Responses to greetings or compliments, and opening or closing remarks.
	PRE-SUGGESTION	Proactive suggestions without a prior user request, or asking whether additional items are needed.
Relation extraction	ATTRIBUTE	A head fashion item has the property given by the tail entity.
	UP_DOWN	The head top item and the tail bottom item form a matched outfit/set.
	PART_OF	The head fashion item is a part of the tail item (sleeve, collar, button, etc.).

Table 1: Labels characterize utterance properties in multimodal commercial dialogues. The table lists labels for intent, strategy, and the three relation extraction predicates; the remaining six tasks use self-explanatory or task-specific output spaces.

closest in domain (multimodal e-commerce agent) but is Chinese and reports a single end-to-end pass[3] metric without per-competence decomposition.

3 Definition

We define below the nine tasks that KoMMERCE evaluates, starting with the multimodal dialogue setting they share. Appendix B provides prompt and output examples for each task.

Dialogue and utterance. A dialogue is a finite sequence $\mathcal{U} = (u_1, \ldots, u_T)$ of utterances produced by either a user or a system, where each utterance $u_t = (x_t, i_t)$ consists of text x_t and an image i_t from an image pool $\mathcal{I}$ (possibly empty). The dialogue prefix is $\mathcal{U}_{\le t} = (u_1, \ldots, u_t)$.

Utterance-level classification (sentiment, intent, strategy). Each task selects a label $\widehat{y}_t^{\tau}$ from a task-specific label set $\mathcal{Y}_\tau$ (Table 1) given the dialogue prefix:

$$\widehat{y}_t^{\tau} = \arg\max_{y \in \mathcal{Y}_\tau} f_\tau(y \mid \mathcal{U}_{\le t}),$$
$$\tau \in \{\text{sent}, \text{intent}, \text{strat}\}.$$

Sentiment ($\mathcal{Y}_{\text{sent}} = \{\text{positive}, \text{negative}, \text{neutral}\}$) predicts the polarity of the user's latest utterance u_t toward an item or its attributes, letting the agent adjust subsequent responses (Appendix A). *Intent* predicts the user's underlying purpose at u_t ($\mathcal{Y}_{\text{intent}}$ in Table 1), gating which downstream action the system should plan. *Strategy* predicts the response strategy for the system's upcoming turn ($\mathcal{Y}_{\text{strat}}$ in Table 1), enabling explicit control over the agent's response form.

Retrieval (image, text knowledge). Each task returns the top-k ranked items $\widehat{R}_\tau$ from a database $\mathcal{D}_\tau$ scored against the dialogue prefix:

$$\widehat{R}_\tau = \underset{d \in \mathcal{D}_\tau}{\text{TopK}}\ s_\tau(d \mid \mathcal{U}_{\le t}), \quad \tau \in \{\text{img}, \text{txt}\}.$$

Image retrieval ($\mathcal{D}_{\text{img}} = \mathcal{I}$, the K-Fashion catalog) supplies candidate items for the agent's recommendation. *Text knowledge retrieval* ($\mathcal{D}_{\text{txt}} = \mathcal{D}$, the curated Korean knowledge base) anchors the agent's textual claims in publicly accessible sources (Appendix A).

State tracking. State tracking predicts the user's accumulated preferences and constraints at turn t as a set of fashion attributes $\widehat{s}_t$ from the dialogue prefix $\mathcal{U}_{\le t}$:

$$\widehat{s}_t = f_{\text{state}}(\mathcal{U}_{\le t}).$$

Accurate state tracking progressively constrains the retrieval search space and ensures the system tracks the user's requirements correctly (Appendix A).

Span detection. Given utterance tokens $x_t = (w_{t,1}, \ldots, w_{t,n_t})$ and image i_t, span detection predicts text spans $\widehat{\mathcal{S}}_t^{\text{text}}$ (contiguous substrings of x_t) and image spans $\widehat{\mathcal{S}}_t^{\text{image}}$ (bounding boxes in i_t) marking fashion-relevant regions:

$$\widehat{\mathcal{S}}_t^{\text{text}} = f_{\text{span}}^{\text{text}}(x_t), \quad \widehat{\mathcal{S}}_t^{\text{image}} = f_{\text{span}}^{\text{image}}(i_t).$$

Spans localize the textual and visual atoms that downstream coreference and relation extraction operate over (Appendix A).

Coreference resolution. Coreference resolution partitions the set of all mentions $\mathcal{M}$ across the dialogue into clusters $\widehat{\Pi}$, where each cluster contains all mentions of one entity:

$$\mathcal{M} = \bigcup_t (\mathcal{S}_t^{\text{text}} \cup \mathcal{S}_t^{\text{image}}),$$
$$\widehat{\Pi} = \{C_1, \ldots, C_J\} \text{ is a partition of } \mathcal{M}.$$

Stable cross-turn coreference is required for coherent entity tracking across multi-turn retrieval and comparison (Appendix A).

Cross-modal relation extraction. Cross-modal relation extraction returns a set of triples $\widehat{\mathcal{R}}$ where each triple links a head entity e_h and a tail entity e_t via a typed predicate $p \in \mathcal{P}$ (Table 1):

$$\widehat{\mathcal{R}} = \{(e_h, p, e_t)\}, \quad p \in \mathcal{P}.$$

To our knowledge, this is the first formulation that pairs entity-level cross-modal RE with multi-turn dialogue grounding: prior cross-modal RE benchmarks (Zheng et al., 2021; He et al., 2023; Huang et al., 2025) are sentence-level, and multimodal dialogue corpora annotate relations only implicitly. Dialogue-level relations also shift with intent and opinion across turns (Gao et al., 2021) (Appendix A).

4 Method

We constructed KoMMERCE on a custom integrated platform that supports multimodal dialogue generation, annotation, and quality review for 100 trained native-Korean crowd workers recruited through a professional data-collection vendor over a three-month period. The platform provides role-based UIs for dialogue speakers, annotators, and reviewers, with built-in retrieval tools, span-marking widgets for both text and image regions, and a session-resumption queue that lets workers continue dialogues started by other workers. Figure 3 shows a representative annotated dialogue. The collection workflow proceeded in three stages followed by a two-level quality review (Figure 2).

4.1 Data Collection

Stage 1. Dialogue generation (Figure 2a. Participants are auto-matched into user-system pairs given a topic sampled from a pool of 15,285 distinct fashion goals. The topic specifies the user persona, target item category, and high-level constraints. Both roles then produce open-ended, free-form turns. Collection screen is depicted in Figure 16).

In this stage each speaker self-annotates *intent*, *sentiment*, and *strategy* in-flow, since these labels are most accurately recovered while the speaker still holds the conversational intent in mind. For each system turn, the worker queries a built-in text retriever and a K-Fashion (AI-Hub, 2020) image retriever and tags the retrieved passages and images used in that turn; these per-turn tags become the gold retrieval annotations against which models are evaluated. Dialogues span multiple sessions: users may leave and later return to continue the same goal, with an average of 2.01 sessions per dialogue, supporting longitudinal evaluation of an agent's session-resumption ability.

Stage 2-1. State-tracking annotation (Figure 2b). A second-pass annotator (distinct from the original dialogue speakers, to prevent self-confirmation bias) labels per-utterance *query state*, the cumulative user constraints encoded as `polarity + (item:attribute-category:attribute-value)` over the K-Fashion taxonomy of 1,992 attribute values, and *clarification state*, system utterances that explicitly request user clarification. For example, a user asking for a beige blouse contributes the query-state element `(top:category:blouse)` and `(top:color:beige)`; a later utterance "but not red" would add the negated element `!(top:color:red)`; and an under-specified utterance such as "something for the top" would contribute the partial element `(top: : )`.

The query state is updated immediately whenever a user utterance adds, replaces, or invalidates a constraint, yielding a per-turn snapshot used as gold for DST evaluation. The leading `!` prefix supports invalidation operations (e.g., a user later rejecting a previously requested color), and unspecified components are left empty within the triple, preserving the exact granularity at which the user expressed each constraint. Detailed Stage 2-1 guidelines are in Appendix G.

Stage 2-2. Cross-utterance annotation (Figure 2c). A third-pass annotator marks *spans* (text substrings or image bounding boxes) on every utterance, groups coreferent spans into entities, and assigns typed predicates between entity pairs to pro-

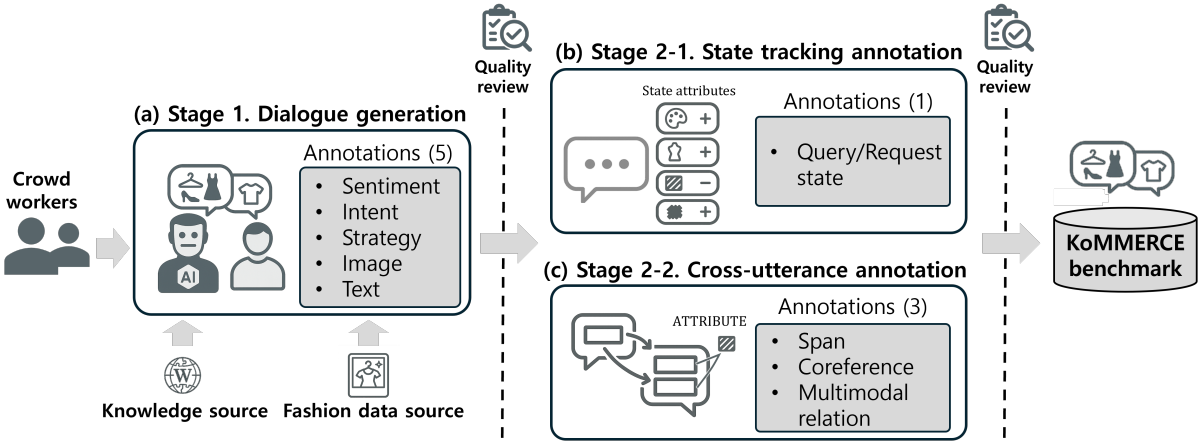


Figure 2: Three-stage data collection pipeline for KoMMERCE. Stage 1 collects dyadic dialogues grounded in fashion images and curated knowledge. Stage 2-1 adds state-tracking labels per utterance. Stage 2-2 annotates cross-utterance spans, coreference clusters, and relation triples.



Figure 3: Annotated dialogue outline. Text/image spans are color-coded with entity IDs; intent in blue; sentiment in dark gray; states in light yellow; relations in the right panel.

duce relation triples. Span boundaries include the head together with the modifiers needed to uniquely specify the referent (e.g., "beige linen dress" rather than just "dress"); when candidate spans overlap, the most informative valid span is kept.

Coreference is annotated across text and image mentions and across sessions, with the explicit goal of producing entity clusters that downstream relation extraction can quantify over. *Relation* predicates are drawn from a fixed set (ATTRIBUTE, UP_DOWN, PART_OF; Table 1) and are annotated at the entity level so each predicate carries from one entity cluster to another rather than between individual mentions; this entity-level annotation produces the cross-modal triples (text-image, image-text) that distinguish KoMMERCE from prior dialogue relation corpora. Detailed Stage 2-2 guidelines are in Appendix G.

Korean-specific annotation. Two properties of Korean shaped the span guidelines. (i) Korean is agglutinative: case particles and verb endings carry grammatical role rather than lexical meaning, so they are excluded from span boundaries to keep spans aligned with the K-Fashion attribute vocabulary. (ii) Korean writes noun-noun compounds without internal spacing, so multi-noun compounds within one orthographic word are tagged as a single span rather than split, matching how a Korean reader parses them.

4.2 Data Quality Control

Quality is enforced along three axes: worker training, source control, and two-level review (Figure 4). **Worker training**: guidelines were iteratively revised three times in consultation with the vendor, validated on a 20-worker pilot, then scaled to 100 workers trained over three months; high-performing workers were promoted to reviewers. **Source control**: images are restricted to license-cleared K-Fashion content with personal identifiers cropped or blurred, and knowledge sources are restricted to the publicly accessible Wikipedia (Wikipedia contributors, 2024) and NamuWiki (NamuWiki contributors, 2024) to exclude untraceable web content. **Two-level review**:

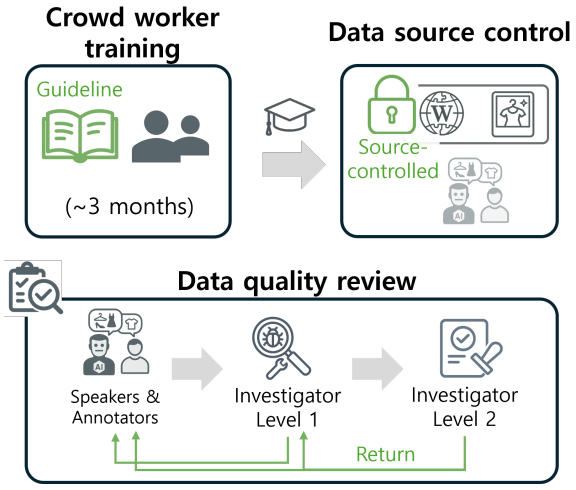


Figure 4: Data quality control workflow: crowd worker training, data source control, and a two-level review.

at every stage, samples must pass a Level 1 review (fit to stage-specific guidelines, with direct correction or return for revision) and a Level 2 second pass before being locked, preventing downstream inconsistencies when subsequent annotation stages are added.

Our sequential pipeline assigns one primary annotator per stage and replaces parallel double-annotation with the L1/L2 review. Roughly 50% of submitted samples passed L1 on the first attempt, with the remainder revised until both L1 and L2 cleared them. To quantify annotation reliability through inter-annotator agreement (IAA), we ran an independent double-annotation study: two annotators independently re-labeled the same 240 selected samples. Appendix C.4 reports IAA and the human baseline.

5 Descriptive Statistics

Before the experiments, we summarize two properties of the KoMMERCE corpus that matter for interpreting it as a benchmark: its scale and diversity, and the structure of the proposed cross-modal relation annotation. Per-task distributions and figures are deferred to Appendix H.

Scale and diversity. KoMMERCE comprises 15,293 dialogues spanning 539,137 utterances, with an average of 35.25 utterances per dialogue and a maximum of 83. Each dialogue is segmented into sessions by temporal gaps, averaging 2.01 sessions per dialogue and supporting longitudinal evaluation that scripted single-session collection rarely provides. The corpus draws on 15,285 unique topics and 29,829 unique fashion images annotated with 686 distinct `item:category:value` combinations in the K-Fashion taxonomy (AI-Hub, 2020)[2], with 1.55 spans per image on average. The full attribute hierarchy is shown in Appendix Figure 17. Within-domain image diversity includes 21 colors, 25 materials, and 21 prints (Table 28).

	# relation	# mention pair					
predicate		All	T-T>0	T-T=0	T-I	I-T	I-I
ATTRIBUTE	91,220	579,447	445,607	66,636	36	67,167	1
UP_DOWN	53,916	243,981	187,807	2,085	54,071	17	1
PART_OF	737	3,184	2,400	65	709	10	0
Sum	145,873	826,612	635,814	68,786	54,816	67,194	2

Table 2: Relations per predicate and mention pairs

Cross-modal relation structure. Table 2 summarizes the annotated cross-modal relations. KoMMERCE contains 145,873 relations across three predicates, dominated by ATTRIBUTE (91,220) and UP_DOWN (53,916). Mention pairs are dominated by cross-utterance text-text relations (T-T>0: 635,814), reflecting the dialogue-level nature of the annotation. Cross-modal pairs (T-I and I-T) are non-trivial as well, arising when textual attributes and top–bottom set relations are grounded to image mentions (Table 2). The utterance-index distance between head and tail mentions has a median of 3 and a 95th percentile of 17 (Appendix Figure 21), indicating that the task tests relation extraction over substantive dialogue spans.

6 Experiments

6.1 Setup

Evaluation conditions. All nine KoMMERCE tasks (Section 3) are run under text-only and multimodal settings at three context-window sizes $k \in \{4, 8, 16\}$, yielding $9 \times 2 \times 3 = 54$ configurations. Full results for each context length and modality are in Appendix D. Appendix C.1 explains how we select evaluation turns or dialogue windows and distinguish text-only from multimodal samples, for example by checking whether their dialogue history contains images.

Execution and reporting. For each configuration we evaluate on a balanced pool of N=250 samples (smaller multimodal sub-pools marked *). Appendix C.2 reports a sample-size sweep supporting this choice for five tasks. Each (model,

[2]The full K-Fashion taxonomy contains 1,992 attribute combinations.

configuration) cell is run three times and we report mean±std. Full sampling, distractor budgets, JSON parsing, and per-task pool details are in Appendix C.

6.2 Baselines and Metrics

Baselines. We benchmark four model families spanning the current proprietary–open-weight frontier. *Anthropic Claude*: Claude 4.5 Sonnet (Anthropic, 2025). *OpenAI GPT*: GPT-5.4 and the smaller GPT-5.4-nano (OpenAI, 2026). *Google Gemini*: Gemini 3.1 Pro and the smaller Gemini 3 Flash (Google DeepMind, 2026). *Alibaba Qwen*: three open-weight variants of Qwen3-VL-32B (Bai et al., 2025), namely [Qwen-Inst] (instruct-tuned reference), [Qwen-Thinking] (thinking enabled, no decoding budget), and [Qwen-Thinking-Budget] (thinking enabled with budget-forced decoding: 3,072 tokens for chain-of-thought and 1,024 tokens reserved for the final answer, exhausting the same 4,096-token decoding cap fairly between thinking and answer). Proprietary baselines use API-default reasoning settings; the client settings and output limits are detailed in Appendix C. All models share the same prompt template. We also compare Qwen3-VL-2B-Instruct before and after task-specific supervised training with low-rank adaptation (LoRA) on the three classification tasks (Section 6.5).

Primary metrics. We report one primary metric per task: Macro-F1 for sentiment, intent, and strategy classification; joint-goal accuracy for state tracking; span-overlap F1 for text spans and bounding-box F1 for image regions in span detection; exact-cluster F1 for coreference resolution (a predicted cluster is counted as a true positive only when its mention set exactly matches a gold cluster's, and per-sample F1 is averaged across samples); triple-level F1 for relation extraction; and Recall@1 for image and text knowledge retrieval. Full metric definitions and the canonical column legend used in the baseline tables are in Appendix A (Table 8). Both retrieval tasks require selecting one answer from 11 candidates (Appendix C.3). For coreference, we additionally report B^3 and LEA to assess partially correct clusters (Appendix D.3).

6.3 Main Results

Full per-model per-task baseline scores are in Appendix D (Tables 13, 14); from those same numbers we highlight two patterns: no model wins more than a few tasks, and every model family scores within the same per-task range.

Task	Winner (Ctx)	Score	Δ_{2nd}
Per-turn prediction tasks			
Sentiment	GPT-5.4 (c=4)	0.593	+0.004†
Intent	Claude-4.5-Sonnet (c=4)	0.975	+0.014
Strategy	Claude-4.5-Sonnet (c=16)	0.949	+0.001†
DST	GPT-5.4 (c=16)	0.876	+0.040
ImgRet	Gemini-3.1-Pro (c=8)	0.918	+0.014
TxtRet	GPT-5.4 (c=16)	0.677	+0.003†
Cross-turn extraction tasks			
Span	Gemini-3-Flash (c=16)	0.795	+0.108
Coref*	GPT-5.4 (c=4)	0.506	+0.012
Relation*	Gemini-3.1-Pro (c=4)	0.942	+0.026

Table 3: Per-task winners across all (model, context) cells. Δ_{2nd} is the margin over the second-best cell. Winners span four model families. †Margin within typical per-run std (< 0.01); should be read as a statistical tie between the listed winner and the runner-up rather than a clear win. *small-pool ctx=4 split. Full numbers in Appendix D.

No single model wins all nine tasks. The nine task winners spread across four model families (Table 3): GPT-5.4 takes a plurality (4/9), Claude 4.5 Sonnet and Gemini 3.1 Pro each lead 2, and Gemini 3 Flash leads 1; the remaining 5/9 still go to a non-GPT family. Three margins (marked † in Table 3) fall within per-run std and should be read as ties. Open-weight Qwen-Inst sits within 0.04 of the best closed-weight model on saturated classification but trails by 0.20 or more on the structured and state tasks.

Every model scores in the same per-task range. All eight model-variants score above 0.88 on Intent and Strategy classification, fall between 0.40 and 0.92 on DST, Span, and Image Retrieval, and stay below 0.60 on Sentiment, Coreference, and Relation Extraction (Appendix D). The pattern is uniform across Anthropic, Google, OpenAI, and Alibaba families, suggesting these splits reflect a shared architectural limitation rather than a single-model artifact, which we examine next.

6.4 Analysis

Context length: candidate validation vs. structured graph construction. The nine tasks split into two groups by how their scores respond to context length (Figure 5). *Candidate-validation tasks* (DST, Span, Image Retrieval) ask the model to se-

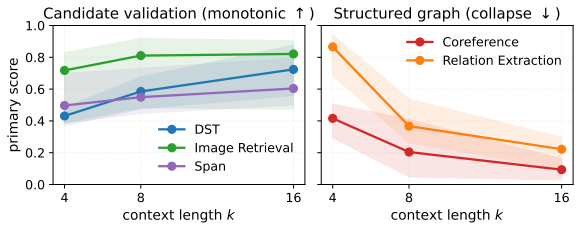


Figure 5: Mean task-specific score across seven model-variants[3] as context length grows; shaded bands span the min–max range. Candidate-validation tasks (left) rise monotonically; structured-graph tasks (right) collapse. Full numbers in Appendix D.

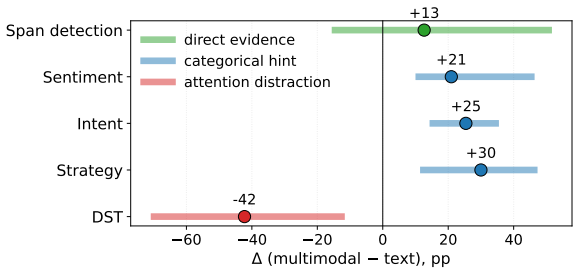


Figure 6: Image-context delta (Δ multimodal − text, in percentage points) across all (model, context) cells. Bars span the min–max range; the dot marks the median. Full per-cell numbers in Appendix D.2.

lect valid items from a closed candidate set; scores rise monotonically with context across all evaluated model-variants (Figure 5, left). *Structured-graph tasks* (Coreference, Relation) instead require linking referents across the entire dialogue, and the number of candidate mention-pairs or triples grows quadratically with context; without an explicit symbol registry, the model must scan an expanding combinatorial space within its attention budget, and performance collapses sharply (Figure 5, right). The practical implication for serving stacks (e.g., vLLM) is that context length should be tuned per task: long context for candidate-validation tasks (DST, retrieval, span), short context for structured-graph tasks (coreference, relation). Enlarging the context window alone does not yield the structured-prediction capability that dialogue-level reference resolution requires.

Image context plays three task-conditional roles. The effect of adding images to the dialogue history depends on what the task predicts (Figure 6). When the target is itself visual (*Span* detection), image is *direct evidence* and lifts performance sharply. When the target is a short categorical label (*Sentiment*, *Intent*, *Strategy*), image is a *categorical hint* that disambiguates the latent dimension. When the target is structured but linguistic (*DST*), image becomes *attention-budget distraction*: DST drops in the multimodal split for most models. Practical implication: image input should be fed to LLM-agent pipelines selectively, always for visual-target tasks (span detection), often for short categorical predictions, sparingly for structured linguistic tasks (DST).

Inference protocol drives reliability more than model identity. Inference protocol can shift a model's reliability profile by an order of magnitude (Table 5). On the same Qwen weights, switching from unconstrained thinking to budget-forced decoding drops empty, truncation, and JSON-fail rates from 24.5%/16.2%/17.9% to 2.8%/0%/1.8%, making the budget-forced configuration more JSON-reliable than several frontier closed-weight baselines. Controlling for inference protocol is therefore essential when comparing model reliability.

Qwen3-VL-32B-Instruct vs. closed-weight: gap is task-heterogeneous. The gap depends on what is being predicted (Table 6): Qwen-Inst sits within 0.04 of the best closed-weight on saturated classification (Intent, Strategy) but lags by 0.22 or more on DST and Span and by 3.3× on Coreference at ctx=16. Conclusions here apply to this single open-weight reference; subject to that, Qwen-Inst is a viable substitute for closed-weight frontier models on simple classification but not yet on dialogue-level structured reasoning.

6.5 Supervised baselines

We train task-specific baselines to assess how supervised adaptation changes classification performance in KoMMERCE. Starting from Qwen3-VL-2B-Instruct, we train a separate LoRA (Hu et al., 2022) adapter for sentiment, intent, and strategy classification, using 5,000 training and 500 validation examples per task. We compare each adapter with the unadapted model on the same k=16 evaluation samples (N=250 per task and setting). Appendix D.6 details the training and evaluation protocol.

Task-specific supervision improves Macro-F1 in all six settings (Table 4). The largest gain is in text-only strategy classification, which increases from 0.235 to 1.000. These results show that KoMMERCE annotations support effective adaptation

Training	Sentiment		Intent		Strategy	
	Text-only	Multimodal	Text-only	Multimodal	Text-only	Multimodal
Zero-shot	0.307	0.437	0.349	0.709	0.235	0.573
Task-specific LoRA	0.480	0.986	0.795	0.991	1.000	0.987

Table 4: Qwen3-VL-2B-Instruct with task-specific supervision (Macro-F1, k=16, N=250 per setting). LoRA uses one training run per task.

Configuration	Empty	Trunc.	JSON-fail
Qwen-Thinking	24.5%	16.2%	17.9%
Qwen-Thinking-Budget	2.8%	0.0%	1.8%

Table 5: Inference-protocol effect on the same Qwen3-VL-32B-Thinking weights (*-Budget* = forced decoding budget). Per-model breakdowns: Appendix Table 23.

Task (ctx=16)	Qwen-Inst	Best closed	Δ
Saturated classification (gap < 0.04)			
Intent	0.926	0.961	−0.035
Strategy	0.920	0.949	−0.029
Structured / state tasks (gap > 0.20)			
DST	0.639	0.876	−0.237
Span	0.572	0.795	−0.223
Coref	0.033	0.108	3.3×

Table 6: Qwen-Inst vs. best closed-weight at ctx=16. Best closed per task: Intent and Span → Gemini-3-Flash; Strategy → Claude-4.5-Sonnet; DST and Coref → GPT-5.4. Full numbers: Appendix D.

of a small multimodal model on the three classification tasks.

7 Conclusion

We introduce KoMMERCE, a Korean multimodal benchmark with nine coordinated NLU tasks, including the first formulation of cross-modal dialogue relation extraction. Baselines reveal failure patterns hidden by end-to-end scores, giving practitioners per-competence signals for deploying and improving commercial LLM agents.

Limitations

Role-play vs. production-log data. Crowd-worker role-play does not replicate the statistical distribution of production logs, which remain valuable complementary evidence. We designed KoMMERCE to mitigate the three artifacts most consistently reported for this collection methodology, but each mitigation reduces rather than eliminates the underlying gap. **(i)** Lexical leakage from task instructions (Gururangan et al., 2018; Geva et al., 2019): the first user utterance shares only 1.77% of its words with the topic instruction on average (median 0), and full verbatim copying occurs in 0.01% of dialogues. **(ii)** Short dialogues from crowd-worker WoZ (Huang et al., 2022; Chen et al., 2020; Zhao et al., 2022): KoMMERCE averages 35.25 utterances per dialogue versus the 8–20 utterances/turns typical of prior corpora. **(iii)** Single-session focus (Xu et al., 2022; Jang et al., 2023): 67% of KoMMERCE dialogues span multiple sessions. These mitigations make KoMMERCE behaviorally non-trivial but do not substitute for real-traffic evaluation.

Sandboxed evaluation. KoMMERCE evaluates dialogue understanding inside a closed corpus, with no live account linkage, no payment execution, and no integration with a production retail backend. Failure modes that surface only in deployment, including request latency under concurrent load, network partitions during transactions, recovery from invalid payment or inventory states, account-level abuse and fraud vectors, and the long tail of system-side errors that downstream agents must compensate for, are therefore outside the benchmark's scope. A model that scores well on KoMMERCE has been shown to parse intent, ground references, and produce structured outputs reliably, but its production safety profile must be re-evaluated against the live system it is deployed into. We recommend end-to-end stress testing against a sandboxed retail backend, or a small-scale shadow deployment on real traffic, as a natural complement to KoMMERCE-style competence audits.

Temporal drift. Fashion vocabulary, popular styles, brand inventory, and the external knowledge sources our retrieval relies on (Wikipedia, NamuWiki) all evolve continuously, so a corpus collected over a fixed three-month window inevitably ages. Retrieval and shallow classification tasks are most exposed: a model evaluated on KoMMERCE may degrade on queries about new

collections, recently launched brands, or trending style terms that never appear in the corpus. Structured tasks (DST, coreference, relation extraction) are comparatively more transferable because their schemas and predicate vocabularies are time-invariant, but the underlying attribute set still drifts whenever the K-Fashion taxonomy is extended. Practitioners deploying KoMMERCE-trained or KoMMERCE-evaluated systems should monitor query-distribution drift after deployment, refresh the benchmark periodically, and where feasible reserve a held-out time slice for temporal-generalization probes; we plan to release versioned updates as part of the dataset distribution.

Ethics Statement

We collect data with informed consent and verify usage rights and provenance for all assets. Personally identifiable information is not included during collection, and we exclude sensitive financial details while applying procedural safeguards. Annotators receive training, fair compensation, and documented guidance with a clear process for resolving disagreements. To address bias and fairness, we exclude content that reinforces stereotypes or evaluates physical appearance. For safety and misuse prevention, the benchmark does not connect to real accounts or enable payment actions. The dataset release provides image URLs and attribute annotations; access to the source images follows the K-Fashion dataset's terms.

Use of AI Assistants During the preparation of this work the author(s) used ChatGPT in order to check for typographical errors and improve the naturalness of the language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Acknowledgement

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220320, Artificial intelligence research about cross-modal dialogue modeling for one-on-one multi-modal interactions).

References

AI-Hub. 2020. K-fashion images. https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=51. Build year 2020; last updated 2021-09-23; accessed 2025-10-22.

Amazon. 2024. Amazon announces Rufus, a new generative AI-powered conversational shopping experience. About Amazon News. Official announcement of the Rufus conversational shopping assistant for recommendations.

Anthropic. 2025. Claude Sonnet 4.5 system card. Anthropic Technical Report. Frontier multimodal LLM evaluated as a closed-weight baseline.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466, Marseille, France. European Language Resources Association.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2023. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126.

Gartner, Inc. 2025. Gartner predicts 40% of enterprise apps will feature task-specific AI agents by 2026, up from less than 5% in 2025. Gartner Press Release. Industry forecast on the rapid uptake of task-specific AI agents in enterprise applications.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of EMNLP-IJCNLP 2019*, pages 1161–1166.

Google DeepMind. 2026. Gemini 3.1 Pro model card. Google DeepMind Model Card. Closed-weight multimodal LLM family used as baselines; the Gemini 3 Flash variant (December 2025) is also evaluated.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL-HLT 2018, Volume 2 (Short Papers)*, pages 107–112.

Liang He, Hongke Wang, Yongchang Cao, Zhen Wu, Jianbing Zhang, and Xinyu Dai. 2023. MORE: A multimodal object-entity relation extraction dataset with a benchmark evaluation. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pages 4564–4573, New York, NY, USA. Association for Computing Machinery.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Seokhee Hong, Sunkyoung Kim, Guijin Son, Soyeon Kim, Yeonjung Hong, and Jinsik Lee. 2025. From KMMLU-Redux to KMMLU-Pro: A professional Korean benchmark suite for LLM evaluation. *Preprint*, arXiv:2507.08924.

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*.

Shizhou Huang, Bo Xu, Changqun Li, Yang Yu, and Xin Alex Lin. 2025. MRE-MI: A multi-image dataset for multimodal relation extraction in social media posts. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Yi Huang, Xiaoting Wu, Si Chen, Wei Hu, Qing Zhu, Junlan Feng, Chao Deng, Zhijian Ou, and Jiangjiang Zhao. 2022. CMCC: A comprehensive and large-scale human-human dataset for dialogue systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 48–61, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of EMNLP 2023*, pages 13584–13606.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Lizi Liao, Le Hong Long, Zheng Zhang, Minlie Huang, and Tat-Seng Chua. 2021. MMConv: An environment for multimodal conversational search across multiple domains. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Introduces RefCOCOg.

Microsoft. 2023. Announcing microsoft copilot, your everyday AI companion. Microsoft Official Blog. Describes Copilot integrating web and work context for AI-powered search and assistance.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

NamuWiki contributors. 2024. Namuwiki. `https://namu.wiki/`. Accessed 2025-10-21, language=ko.

OpenAI. 2025. Buy it in ChatGPT: Instant checkout and the agentic commerce protocol. OpenAI Blog. Introduces Instant Checkout and the Agentic Commerce Protocol enabling in-chat payments.

OpenAI. 2026. GPT-5.4 and GPT-5.4-nano. OpenAI Model Documentation. Closed-weight multimodal LLM family used as baselines.

Kaicheng Pang, Xingxing Zou, and Waikeung Wong. 2025. FashionM3: Multimodal, multitask, and multiround fashion assistant based on unified vision–language model. *arXiv preprint arXiv:2504.17826*.

Perplexity AI. 2025. Introducing the perplexity search API. Perplexity Blog. Official announcement describing Perplexity's answer engine infrastructure and developer API.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 143–152.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Amrita Saha, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Jaehyung Seo, Jaewook Lee, Chanjun Park, SeongTae Hong, Seungjun Lee, and Heuiseok Lim. 2024. KoCommonGEN v2: A benchmark for navigating Korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2390–2415.

Hyopil Shin, Sangah Lee, Dongjun Jang, Wooseok Song, Jaeyoon Kim, Chaeyoung Oh, Hyemi Jo, Youngchae Ahn, Sihyun Oh, Hyohyeong Chang, Sunkyoung Kim, and Jinsik Lee. 2025. KoBALT: Korean benchmark for advanced linguistic tasks. *Preprint*, arXiv:2505.16125.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. KMMLU: Measuring massive multitask language understanding in Korean. *Preprint*, arXiv:2402.11548.

Statista Market Insights. 2025. eCommerce – south Korea. Statista Market Outlook. Market sizing for the South Korean e-commerce sector; accessed 2026-05-26.

Stripe. 2025. Developing an open standard for agentic commerce. Stripe Blog. Details ACP and how payments can be executed within AI agents, including ChatGPT.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Haoxin Wang, Xianhan Peng, Xucheng Huang, Yizhe Huang, Ming Gong, Chenghan Yang, Yang Liu, and Ling Jiang. 2025a. ECom-Bench: Can LLM agent resolve real-world E-commerce customer support issues? *Preprint*, arXiv:2507.05639.

Zihan Wang, Xiaocui Yang, YongKang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025b. MUSE: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0. LDC2013T19, Linguistic Data Consortium.

Wikipedia contributors. 2024. Wikipedia. `https://www.wikipedia.org/`. Accessed 2025-10-21.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. 2023. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291, Toronto, Canada. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of ACL 2022 (Volume 1: Long Papers)*, pages 5180–5197.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In *Proceedings of the European Conference on Computer Vision*. Introduces RefCOCO and RefCOCO+.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Nan Zhao, Haoran Li, Youzheng Wu, and Xiaodong He. 2022. JDDC 2.1: A multimodal Chinese dialogue dataset with joint tasks of query rewriting, response generation, discourse parsing, and summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12037–12051, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021. MNRE: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *IEEE International Conference on Multimedia and Expo (ICME)*.

A Evaluation Metrics

Each benchmark instance is converted into a task-specific prompt, and the model is required to return a structured JSON response. We report one primary metric per task, matching the benchmark tables. For readability, the examples below are shown as gray prompt/output boxes.

Classification (sentiment, intent, strategy): Macro-F1. Let $\{(y_i, \hat{y}_i)\}_{i=1}^{N}$ denote the gold and predicted labels after converting both labels to uppercase strings, and let

$$\mathcal{Y} = \{\hat{y}_i \mid 1 \le i \le N\} \cup \{y_i \mid 1 \le i \le N\}. \quad (1)$$

For each $c \in \mathcal{Y}$,

$$\mathrm{TP}_c = \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = c \wedge y_i = c], \quad (2)$$

$$\mathrm{FP}_c = \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = c \wedge y_i \neq c], \quad (3)$$

$$\mathrm{FN}_c = \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i \neq c \wedge y_i = c], \quad (4)$$

and

$$P_c = \begin{cases} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} & \text{if } \mathrm{TP}_c + \mathrm{FP}_c > 0, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

$$R_c = \begin{cases} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} & \text{if } \mathrm{TP}_c + \mathrm{FN}_c > 0, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

$$F1_c = \begin{cases} \frac{2P_cR_c}{P_c + R_c} & \text{if } P_c + R_c > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

We report

$$\text{Macro-F1} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} F1_c. \quad (8)$$

DST: joint-goal accuracy. Let $\hat{Q}_i$ and Q_i denote the predicted and gold `query_state` sets after keeping only non-empty strings and applying `strip()` to each entry; `request_state` is not scored. We report

$$\mathrm{JGA} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{Q}_i = Q_i]. \quad (9)$$

Structured prediction (span, coreference, relation): F1-overlap. For span detection, coreference, and relation extraction, let $\hat{A}_i$ and A_i be the predicted and gold sets of exact atomic items. The score is zero when exactly one set is empty ($\oplus$). The per-instance score is

$$F1_i(A) = \begin{cases} 1 & \text{if } \hat{A}_i = A_i = \varnothing, \\ 0 & \text{if } \hat{A}_i = \varnothing \oplus A_i = \varnothing, \\ \frac{2P_i(A)R_i(A)}{P_i(A) + R_i(A)} & \text{otherwise,} \end{cases} \quad (10)$$

where

$$P_i(A) = \frac{|\hat{A}_i \cap A_i|}{|\hat{A}_i|}, \qquad R_i(A) = \frac{|\hat{A}_i \cap A_i|}{|A_i|}. \quad (11)$$

The reported score is $\frac{1}{N} \sum_{i=1}^{N} F1_i(A)$.

Retrieval (image, text knowledge): accuracy. For image retrieval and text knowledge retrieval, we evaluate only the final selected candidate index:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{z}_i = z_i]. \quad (12)$$

B Task Evaluation

For each of the nine KoMMERCE tasks we describe the input the model sees, the JSON structure it must return, and the scoring rule applied to that output. Each task block ends with a worked prompt/output example.

Sentiment analysis. Given the dialogue history and the latest user utterance, the model predicts one label in {`positive`, `negative`, `neutral`} and is evaluated by Macro-F1 (Figure 7).

Intent classification. Given the dialogue history and the latest user utterance, the model predicts one intent label (e.g., `REQUEST`, `REFINED REQUEST`, `DETAIL`, `CHAT`) and is evaluated by Macro-F1 (Figure 8).

Strategy classification. Given the dialogue history and the latest system utterance, the model predicts one strategy label (e.g., `SUGGESTION`, `CLARIFICATION`, `DETAIL`) and is evaluated by Macro-F1 (Figure 9).

Dialogue state tracking. Given the dialogue history and a dynamically constructed tag menu, the model predicts the current dialogue state as JSON with `query_state`, `request_state`, and a short summary. The benchmark reports only JGA on `query_state` (Figure 10).

Table 7: Primary metrics reported in KoMMERCE.

Task	Prediction target	Reported metric
Sentiment	sentiment label of u_t	Macro-F1
Intent	intent label of u_t	Macro-F1
Strategy	strategy label of u_t	Macro-F1
DST	`query_state` of s_t	JGA
Span detection	text spans / image boxes	F1-overlap
Coreference resolution	mention clusters Π	F1-coref
Relation extraction	relation triples R	F1-relation
Image retrieval	candidate image index	Accuracy
Text knowledge retrieval	candidate knowledge index	Accuracy

Sentiment analysis

Prompt.
`Analyze the sentiment of the user's last utterance given the dialogue history.`

Input.

User. `I want a beige dress.`

System. `Here is one with a softer silhouette.`

...

User. `This is exactly what I wanted. I love it.`

Output.

```
{
  "sentiment": "positive"
}
```

Figure 7: Sentiment analysis: prompt and output example.

Intent classification

Prompt.
`Classify the intent of the user's last utterance given the dialogue history and label definitions.`

Input.

User. `I am looking for a blouse.`

System. `What style do you prefer?`

...

User. `Could you show me a more casual one?`

Output.

```
{
  "intent": "REFINED REQUEST"
}
```

Figure 8: Intent classification: prompt and output example.

Strategy classification

Prompt.
`Classify the strategy of the system's last utterance given the dialogue history and label definitions.`

Input.

User. `I want a black blouse.`

System. `Let me narrow down the options.`

...

System. `Would you prefer a fitted or oversized silhouette?`

Output.

```
{
  "strategy": "CLARIFICATION"
}
```

Figure 9: Strategy classification: prompt and output example.

Span detection. Given either a dialogue context with candidate text spans or an image with candidate boxes, the model predicts all valid spans/boxes in JSON. Predicted and gold items are matched as true positives when their character-IoU (text spans, computed over character indices within the same turn) or box-IoU (image bounding boxes) is at least 0.5; per-instance precision/recall/F1 are then computed and averaged across samples (Figure 11).

Coreference resolution. Given the dialogue history and a mention pool indexed as `M0`, `M1`, . . ., the model groups mentions that refer to the same entity and returns JSON clusters. Each predicted cluster is treated as one unordered exact-match item, and the benchmark reports F1-coref (Figure 12).

Relation extraction. Given the dialogue history, the mention pool, and a fixed list of candidate `(head, tail)` mention pairs, the model assigns

Dialogue state tracking

Prompt.
Infer the user's current query state from the dialogue history and candidate tags.

Input.

User. I need a blouse for work.

System. Do you have a preferred color?

...

User. Preferably white and not too formal.

Candidate tags. top:category:blouse, top:category:dress, top:color:white, top:color:black, top:style:formal

Output.

```
{
  "query_state": [
    "top:category:blouse",
    "top:color:white"
  ],
  "request_state": [],
  "summary": "The user wants a white blouse for work."
}
```

Figure 10: Dialogue state tracking: prompt and output example.

Span detection

Prompt.
Select all valid fashion-related spans from the candidate list.

Input.

User. I want a linen dress today.

System. I can help you find one.

Candidates. [0,9,14], [0,15,20], [0,21,26]

Output.

```
{
  "spans": [[0,9,14], [0,15,20]]
}
```

Figure 11: Span detection: prompt and output example.

each pair a relation label or NONE for no relation and returns the labeled triples as JSON. Pairs labeled NONE carry no relation and act as distractors; they are dropped before scoring, while gold pairs answered NONE become false negatives. The remaining triples are evaluated by exact-match F1-overlap, comparing head/tail mention IDs and the

Coreference resolution

Prompt.
Group mentions that refer to the same fashion item.

Input.

User. I like the black dress.



System. This one looks elegant.

Mentions. M0: "the black dress", M1: bbox[0.12,0.31,0.28,0.40], M2: "This one"

Output.

```
{
  "clusters": [["M0", "M1", "M2"]]
}
```

Figure 12: Coreference resolution: prompt and output example.

relation label after lowercasing/stripping and mapping attribute to attr (Figure 13).

Relation extraction

Prompt.
Assign one relation type to each candidate mention pair.

Input.

User. I want a beige top.

System. Pair the top with a skirt.

Mentions. M0: "top", M1: "beige", M2: "skirt"

Candidate pairs. (M0, M1), (M0, M2), (M1, M2)

Output.

```
{
  "relations": [
    ["M0", "M1", "ATTRIBUTE"],
    ["M0", "M2", "UP_DOWN"],
    ["M1", "M2", "NONE"]
  ]
}
```

Figure 13: Relation extraction: prompt and output example.

Image retrieval. Given the dialogue history and a set of candidate images, the model predicts one candidate index in JSON. The reported score is accuracy over the final selected index (Figure 14).

Image retrieval

Prompt.
Select the image that the system should recommend given the dialogue history.

Input.

User. Show me a white blouse for work.

System. Sure, I will find a few options.

Candidates. <image 1>, <image 2>, <image 3>

Output.

```
{
  "index": 2
}
```

Figure 14: Image retrieval: prompt and output example.

Text knowledge retrieval. Given the dialogue history and a set of candidate knowledge entries, the model predicts one candidate index in JSON. The reported score is accuracy over the final selected index (Figure 15).

Text knowledge retrieval

Prompt.
Select the most appropriate knowledge entry given the dialogue history.

Input.

User. What fabric is suitable for summer?

System. Let me check the fabric guide.

Candidates. <entry 1>, <entry 2>, <entry 3>

Output.

```
{
  "index": 1
}
```

Figure 15: Text knowledge retrieval: prompt and output example.

C Evaluation Protocol Details

This appendix collects the implementation details elided from Section 6.1. Pools are fixed by seed (`seed=42`) and shared across all models and reruns. Distractor pools are pre-baked per sample so grading is deterministic and independent of any external lookup at inference time: 15 distractors per state-tracking sample, 5 per span-detection sample, and 10 per image-retrieval or text-knowledge-retrieval sample. Each instance is rendered with a task-specific prompt and the model returns a JSON object that is parsed and graded against the gold annotation; image turns are inlined in-place at `[IMAGE]` markers in the prompt so multimodal models receive the original turn order. Three output-reliability rates (empty predictions, token-truncation, JSON-parse failures) are tracked alongside primary metrics because they directly affect whether a model can be deployed in practice: a high-scoring model that frequently returns empty or unparseable outputs is unusable. Per-model aggregates are in Appendix E.

Reasoning configuration. Table 9 lists the reasoning/thinking setting for each of the eight model variants. To compare budget forcing at fixed weights, Qwen-Thinking and Qwen-Thinking-Budget use the same Thinking checkpoint; Qwen-Inst uses a separate Instruct checkpoint. Proprietary baselines use API-default reasoning settings. Qwen and the OpenAI-compatible clients set a 4,096-token output limit, while the Gemini client leaves this limit unspecified.

C.1 Evaluation samples and selection

Evaluation sample definition. For task τ and context-window size $k \in \{4, 8, 16\}$, measured in dialogue turns, an evaluation sample is

$$(\mathcal{U}_{\leq t}^{(k)},\ y_{t,k}^{\tau}),$$

where $\mathcal{U}_{\leq t}^{(k)}$ is the dialogue context provided from $\mathcal{U}_{\leq t}$ under this window setting, and $y_{t,k}^{\tau}$ is the corresponding gold annotation. For classification, $y_{t,k}^{\tau}$ is a sentiment, intent, or strategy label; for DST, it is a set of query-state tags. For span detection, coreference resolution, and relation extraction, it contains text spans or image bounding boxes, clusters of coreferent mentions, and relation triples, respectively. For image and text-knowledge retrieval, it is the index of the gold candidate in the candidate list.

Text-only vs. multimodal. We evaluate text-only and multimodal settings separately to examine how task performance and difficulty differ when images are involved. We form these pools by selecting dialogue histories or annotated targets, rather than by removing images from multimodal inputs. For classification, DST, and retrieval, text-only samples are selected from turns whose original preceding history contains no images, while multimodal samples have an image in the last k history turns. For

Table 8: Definitions of the canonical result columns used in the baseline tables. Aggregate columns are overall results, not modality-specific subsets unless explicitly marked otherwise.

Column	Definition	Primary metric
Sent.	Sentiment classification on all labeled sentiment turns.	Macro-F1
Intent	Intent classification on all labeled user turns.	Macro-F1
Strat.	System strategy classification on all labeled system turns.	Macro-F1
DST	Dialogue state tracking on all state-labeled samples.	JGA
Span	Overall span detection over all span samples (text spans + image boxes).	F1-overlap
Coref	Overall coreference resolution over all coreference samples.	F1-coref
Rel.	Overall relation extraction over all relation samples.	F1-relation
ImgRet	Overall image retrieval over both text-only and multimodal contexts.	Accuracy
TxtRet	Overall text knowledge retrieval over both text-only and multimodal contexts.	Accuracy

Model variant	Family	Reasoning setting
Gemini-3-Flash	Google	API defaults
Gemini-3.1-Pro	Google	API defaults
Claude-4.5-Sonnet	Anthropic	API defaults
GPT-5.4	OpenAI	API defaults
GPT-5.4-nano	OpenAI	API defaults
Qwen-Inst	Alibaba	non-thinking (instruct)
Qwen-Thinking	Alibaba	thinking on, no budget
Qwen-Thinking-Budget	Alibaba	thinking on, budget-forced (3,072 CoT / 1,024 answer)

Table 9: Client-side reasoning configuration for each model variant. Only the two Qwen Thinking variants share a checkpoint. API defaults indicate endpoint reasoning settings without client-side overrides.

	Text-only	Multimodal		
Task	Each k	$k = 4$	$k = 8$	$k = 16$
Sentiment	250	250	250	250
Intent	250	250	250	250
Strategy	250	250	250	250
DST	250	250	250	250
Span	250	250	250	250
Coreference	250	15	250	250
Relation	250	3	250	250
Image retrieval	250	250	250	250
Text retrieval	250	250	250	250

Table 10: Evaluation sample counts for each task, setting, and context length, before the three reruns. Each text-only setting has 250 samples at each of $k = 4, 8, 16$. At $k = 4$, the multimodal coreference and relation sets contain 15 and 3 samples, respectively; their scores are marked with stars in the baseline tables.

span detection, coreference, and relation extraction, images also form part of the prediction targets.

DST target construction. To evaluate state tracking within the context shown to the model, we construct DST gold targets from query-state updates over that window. We take the query-state tags at the evaluated turn and remove any tags already present in the state immediately before the window. We retain only samples with at least one remaining tag, so each sample requires a state update.

Shared evaluation pools. To compare models on the same evidence, we fix the evaluation pool for each configuration and share it across all models and three reruns. We sample without replacement from eligible samples using random seeds derived from seed 42. Sentiment uses quotas of 183 neutral, 40 positive, and 27 negative samples; other tasks use uniform sampling. When fewer than 250 eligible samples are available, we use all of them. Table 10 reports the resulting pool sizes, including the small multimodal sub-pools marked with stars in Tables 14 and 17.

Sampling across context lengths. To account for images or DST state updates falling outside shorter context windows, we select all multimodal pools and the text-only DST pool separately for each k. The other text-only pools reuse samples across context lengths. Scores across k can therefore reflect differences in both context and sample composition.

C.2 Justification for the $N = 250$ evaluation pool

To assess whether $N = 250$ is sufficient for evaluation, we use a sample-size sweep ($N = 50$–$1{,}600$) on sentiment, intent, strategy, span detection, and coreference. Table 11 reports the same primary metrics as the benchmark, using the first 50, 100, 200, 400, 800, and 1,600 samples of one fixed Qwen prediction run per task. Each larger subset includes the smaller subsets, with model outputs held fixed. Growing the evaluation pool eightfold from 200 to 1,600 changes the scores by at most 0.0333, and by less than 0.01 for intent, span detection, and coreference. These small changes support the choice of $N = 250$ for the five tasks examined.

Task	Metric	$N = 50$	100	200	400	800	1,600	$\lvert\Delta_{200\to1600}\rvert$
Sentiment	Macro-F1	.4619	.5104	.4912	.4977	.5263	.5245	.0333
Intent	Macro-F1	.9451	.8946	.9155	.9008	.9138	.9129	.0026
Strategy	Macro-F1	.8482	.8792	.9065	.9276	.9361	.9378	.0313
Span	Overlap F1	.6086	.6326	.6380	.6410	.6501	.6465	.0085
Coreference	Exact F1	.0969	.0955	.0796	.0770	.0778	.0822	.0026

Table 11: Sample-size sweep to assess the choice of an $N = 250$ evaluation pool. Scores use the first N samples of a fixed Qwen prediction run for each task. The final column gives the absolute difference between $N = 200$ and $N = 1{,}600$, calculated from the displayed four-decimal scores.

C.3 Retrieval candidate construction

Shared candidate format. Both retrieval tasks use single-answer selection over the same number of candidates: a shuffled list of one gold answer and ten distractors. When a turn has multiple annotated image or article references, we use the first as the gold candidate. With one gold answer among the 11 candidates, top-1 accuracy is equivalent to Recall@1.

Image candidates. Image retrieval uses ten random distractor images from the image pool, excluding all gold images annotated for the target turn. Candidate images are shown in both the text-only and multimodal history settings.

Text-knowledge candidates. Each text-knowledge candidate is constructed from passages quoted in the dialogues, rather than the full source article. We group cited system utterances by article ID and display each candidate as a title and a body preview derived from these passages.

To make text-knowledge retrieval less susceptible to surface keyword matching, we use five TF-IDF hard negatives and five random articles as distractors. We select the hard negatives by character-bigram/trigram TF-IDF cosine similarity to the gold article's quoted passages, then sample five other articles uniformly. All articles cited anywhere in the same dialogue are excluded from the distractors.

C.4 Assessing annotation quality

Experimental setup. To assess annotation quality, we ran an independent double-annotation study with two annotators. Both independently re-label the same 30 samples for each combination of four tasks (sentiment, intent, strategy, and coreference) and two settings (text-only and multimodal), totaling 240 samples. Samples are drawn from the model evaluation pools, using the $k = 16$ pools for multimodal evaluation. Annotators see up to 16 preceding turns, the target utterance for classification, and images for multimodal samples. For coreference, we select one entity with at least two mentions per dialogue, provide a text mention as an anchor, and ask annotators which mentions in a supplied candidate list co-refer with it. Gold answers are withheld.

Inter-annotator agreement. IAA measures whether independent annotators make consistent judgments. For sentiment, intent, and strategy, we compute Cohen's κ between the annotators' labels; for coreference, link-κ is Cohen's κ over their per-candidate yes/no linking decisions. Intent and strategy show almost-perfect agreement ($\kappa = 0.815$–0.914), and coreference shows substantial agreement (link-$\kappa = 0.703$–0.721; Table 12). Sentiment agreement is moderate, with $\kappa = 0.507$ for text-only samples and 0.545 for multimodal samples. This label uncertainty should be considered when interpreting model performance on sentiment.

Human performance. Human annotations closely match the gold intent and strategy labels, with mean accuracy of 0.883–0.967 (Table 12). We measure classification accuracy against gold and use link-F1 to measure how accurately annotators recover gold coreference links. For link-F1, decisions are pooled within each setting for each annotator; both accuracy and link-F1 are then averaged over the two annotators.

D Full Per-Model Per-Task Baseline Results

This appendix provides the full results supporting the model comparisons and analyses in Sections 6.3 and 6.4. We report modality-combined and modality-split baseline results, followed by additional coreference and relation analyses, a controlled image ablation, and the protocol for task-specific supervised baselines.

Task	Modality	N	IAA metric	IAA	Human score	Human metric
Sentiment	Text-only	30	Cohen's κ	.507	.650	Accuracy
Sentiment	Multimodal	30	Cohen's κ	.545	.883	Accuracy
Intent	Text-only	30	Cohen's κ	.815	.883	Accuracy
Intent	Multimodal	30	Cohen's κ	.914	.967	Accuracy
Strategy	Text-only	30	Cohen's κ	.816	.883	Accuracy
Strategy	Multimodal	30	Cohen's κ	.871	.933	Accuracy
Coreference	Text-only	30	Link-κ	.703	.460	Link-F1
Coreference	Multimodal	30	Link-κ	.721	.813	Link-F1

Table 12: Inter-annotator agreement (IAA) and human baseline on 30 samples per task–setting combination (240 total), labeled by two annotators. Human scores are averaged over the two annotators.

D.1 Modality-combined results

To compare overall task performance across models and context lengths, we compute each primary metric over the combined text-only and multimodal evaluation samples. Table 13 reports classification, state tracking, and retrieval results; Table 14 reports span detection, coreference resolution, and relation extraction results. Each model–context entry reports the mean and standard deviation over three reruns.

D.2 Modality-split results

To compare performance between text-only and multimodal settings, Tables 15, 16, and 17 report the two settings separately for classification, state tracking and retrieval, and cross-turn extraction, respectively. These results provide the setting-specific scores underlying the image-context analysis in Section 6.4; sample selection is described in Appendix C.1.

Table 13: Primary-metric baseline results with model-context pairs as rows for per-turn prediction tasks. Each cell reports mean ± std over three reruns.

Model	Ctx	Sent.	Intent	Strat.	DST	ImgRet	TxtRet
Gemini-3-Flash	4	0.516 ± 0.002	0.927 ± 0.067	0.924 ± 0.002	0.483 ± 0.126	0.779 ± 0.001	0.652 ± 0.002
Gemini-3-Flash	8	0.434 ± 0.002	0.957 ± 0.006	0.948 ± 0.001	0.663 ± 0.050	0.892 ± 0.009	0.631 ± 0.007
Gemini-3-Flash	16	0.426 ± 0.005	0.961 ± 0.007	0.945 ± 0.003	0.807 ± 0.010	0.904 ± 0.007	0.670 ± 0.008
Gemini-3.1-Pro	4	0.582 ± 0.072	0.932 ± 0.067	0.880 ± 0.065	0.456 ± 0.126	0.829 ± 0.010	0.634 ± 0.005
Gemini-3.1-Pro	8	0.480 ± 0.061	0.922 ± 0.068	0.906 ± 0.065	0.654 ± 0.063	**0.918 ± 0.002**	0.649 ± 0.003
Gemini-3.1-Pro	16	0.509 ± 0.003	0.960 ± 0.000	0.947 ± 0.001	0.836 ± 0.015	0.904 ± 0.031	0.669 ± 0.010
Claude-4.5-Sonnet	4	0.589 ± 0.002	**0.975 ± 0.001**	0.918 ± 0.000	0.449 ± 0.123	0.745 ± 0.004	0.633 ± 0.002
Claude-4.5-Sonnet	8	0.472 ± 0.002	0.953 ± 0.001	0.948 ± 0.000	0.612 ± 0.057	0.821 ± 0.005	0.645 ± 0.004
Claude-4.5-Sonnet	16	0.452 ± 0.003	0.955 ± 0.002	**0.949 ± 0.000**	0.785 ± 0.017	0.855 ± 0.012	0.665 ± 0.007
GPT-5.4	4	**0.593 ± 0.002**	0.904 ± 0.061	0.896 ± 0.003	0.477 ± 0.136	0.769 ± 0.008	0.646 ± 0.003
GPT-5.4	8	0.484 ± 0.002	0.860 ± 0.067	0.938 ± 0.001	0.676 ± 0.072	0.875 ± 0.002	0.658 ± 0.006
GPT-5.4	16	0.465 ± 0.002	0.805 ± 0.002	0.798 ± 0.001	**0.876 ± 0.009**	0.891 ± 0.007	**0.677 ± 0.005**
GPT-5.4-nano	4	0.575 ± 0.005	0.885 ± 0.006	0.809 ± 0.003	0.378 ± 0.123	0.389 ± 0.012	0.601 ± 0.005
GPT-5.4-nano	8	0.459 ± 0.007	0.870 ± 0.011	0.856 ± 0.006	0.476 ± 0.050	0.483 ± 0.004	0.616 ± 0.011
GPT-5.4-nano	16	0.417 ± 0.006	0.866 ± 0.008	0.686 ± 0.012	0.560 ± 0.035	0.474 ± 0.020	0.630 ± 0.003
Qwen-Inst	4	0.536 ± 0.001	0.954 ± 0.001	0.895 ± 0.000	0.402 ± 0.110	0.760 ± 0.003	0.621 ± 0.001
Qwen-Inst	8	0.423 ± 0.001	0.920 ± 0.000	0.932 ± 0.000	0.527 ± 0.051	0.835 ± 0.002	0.636 ± 0.004
Qwen-Inst	16	0.400 ± 0.001	0.926 ± 0.001	0.920 ± 0.000	0.639 ± 0.008	0.847 ± 0.001	0.661 ± 0.001
Qwen-Thinking	4	0.558 ± 0.005	0.923 ± 0.001	0.810 ± 0.005	0.284 ± 0.113	0.719 ± 0.009	0.605 ± 0.005
Qwen-Thinking	8	0.445 ± 0.007	0.810 ± 0.057	0.765 ± 0.060	0.326 ± 0.059	0.833 ± 0.007	0.626 ± 0.005
Qwen-Thinking	16	0.431 ± 0.002	0.762 ± 0.004	0.736 ± 0.056	0.406 ± 0.020	0.875 ± 0.009	0.651 ± 0.001
Qwen-Thinking-Budget	4	0.555 ± 0.002	0.923 ± 0.006	0.809 ± 0.009	0.371 ± 0.138	0.751 ± 0.002	0.624 ± 0.007
Qwen-Thinking-Budget	8	0.447 ± 0.002	0.895 ± 0.003	0.840 ± 0.006	0.488 ± 0.078	0.851 ± 0.007	0.636 ± 0.006
Qwen-Thinking-Budget	16	0.428 ± 0.006	0.890 ± 0.005	0.805 ± 0.002	0.565 ± 0.012	0.873 ± 0.002	0.674 ± 0.011

Table 14: Primary-metric baseline results with model-context pairs as rows for cross-turn extraction tasks. Each cell reports mean ± std over three reruns.

Model	Ctx	Span	Coref	Rel.
Gemini-3-Flash	4	0.698 ± 0.005	0.411 ± 0.004*	0.810 ± 0.005*
Gemini-3-Flash	8	0.730 ± 0.009	0.240 ± 0.004	0.512 ± 0.009
Gemini-3-Flash	16	**0.795 ± 0.006**	0.153 ± 0.002	0.297 ± 0.004
Gemini-3.1-Pro	4	0.588 ± 0.016	0.494 ± 0.011*	**0.942 ± 0.001***
Gemini-3.1-Pro	8	0.640 ± 0.002	0.409 ± 0.006	0.538 ± 0.003
Gemini-3.1-Pro	16	0.687 ± 0.002	0.162 ± 0.005	0.283 ± 0.003
Claude-4.5-Sonnet	4	0.477 ± 0.006	0.405 ± 0.007*	0.913 ± 0.003*
Claude-4.5-Sonnet	8	0.522 ± 0.006	0.155 ± 0.006	0.338 ± 0.010
Claude-4.5-Sonnet	16	0.587 ± 0.007	0.135 ± 0.001	0.215 ± 0.004
GPT-5.4	4	0.467 ± 0.009	**0.506 ± 0.005***	0.916 ± 0.004*
GPT-5.4	8	0.488 ± 0.005	0.282 ± 0.002	0.263 ± 0.005
GPT-5.4	16	0.513 ± 0.007	0.108 ± 0.004	0.214 ± 0.001
GPT-5.4-nano	4	0.387 ± 0.005	0.294 ± 0.005*	0.910 ± 0.005*
GPT-5.4-nano	8	0.447 ± 0.006	0.048 ± 0.003	0.281 ± 0.007
GPT-5.4-nano	16	0.501 ± 0.004	0.033 ± 0.002	0.096 ± 0.003
Qwen-Inst	4	0.425 ± 0.003	0.477 ± 0.002*	0.883 ± 0.002*
Qwen-Inst	8	0.517 ± 0.001	0.106 ± 0.000	0.286 ± 0.002
Qwen-Inst	16	0.572 ± 0.003	0.033 ± 0.001	0.193 ± 0.001
Qwen-Thinking	4	0.087 ± 0.004	0.352 ± 0.023*	0.727 ± 0.008*
Qwen-Thinking	8	0.067 ± 0.009	0.183 ± 0.013	0.241 ± 0.005
Qwen-Thinking	16	0.060 ± 0.006	0.008 ± 0.001	0.045 ± 0.004
Qwen-Thinking-Budget	4	0.436 ± 0.003	0.325 ± 0.010*	0.686 ± 0.016*
Qwen-Thinking-Budget	8	0.501 ± 0.010	0.190 ± 0.006	0.352 ± 0.004
Qwen-Thinking-Budget	16	0.573 ± 0.002	0.029 ± 0.005	0.256 ± 0.006

Table 15: Per-turn classification tasks: modality split performance. Tasks are partitioned based on the presence of images in the dialogue history.

Model	Ctx	Sent-T	Sent-M	Intent-T	Intent-M	Strat-T	Strat-M
Gemini-3-Flash	4	0.329 ± 0.002	0.660 ± 0.004	0.727 ± 0.061	0.993 ± 0.003	0.749 ± 0.077	0.965 ± 0.001
Gemini-3-Flash	8	0.341 ± 0.003	0.508 ± 0.007	0.725 ± 0.060	0.946 ± 0.009	0.695 ± 0.077	**0.991 ± 0.002**
Gemini-3-Flash	16	0.331 ± 0.005	0.501 ± 0.009	0.727 ± 0.062	0.967 ± 0.004	0.695 ± 0.077	0.986 ± 0.001
Gemini-3.1-Pro	4	0.355 ± 0.002	0.792 ± 0.099	0.734 ± 0.063	**0.994 ± 0.000**	**0.803 ± 0.000**	0.917 ± 0.068
Gemini-3.1-Pro	8	0.331 ± 0.043	0.654 ± 0.007	0.731 ± 0.067	0.970 ± 0.005	**0.803 ± 0.000**	0.946 ± 0.068
Gemini-3.1-Pro	16	0.357 ± 0.005	0.625 ± 0.004	**0.778 ± 0.002**	0.954 ± 0.001	**0.803 ± 0.000**	0.985 ± 0.001
Claude-4.5-Sonnet	4	0.347 ± 0.002	**0.811 ± 0.003**	0.776 ± 0.001	0.985 ± 0.000	**0.803 ± 0.000**	0.925 ± 0.000
Claude-4.5-Sonnet	8	0.358 ± 0.002	0.573 ± 0.002	0.776 ± 0.002	0.919 ± 0.004	**0.803 ± 0.000**	0.984 ± 0.000
Claude-4.5-Sonnet	16	0.350 ± 0.004	0.540 ± 0.001	0.776 ± 0.002	0.954 ± 0.006	**0.803 ± 0.000**	0.988 ± 0.000
GPT-5.4	4	0.346 ± 0.004	0.805 ± 0.008	0.625 ± 0.004	0.978 ± 0.001	0.609 ± 0.011	0.906 ± 0.002
GPT-5.4	8	0.362 ± 0.007	0.582 ± 0.006	0.676 ± 0.062	0.911 ± 0.069	0.629 ± 0.001	0.981 ± 0.001
GPT-5.4	16	0.364 ± 0.002	0.539 ± 0.001	0.672 ± 0.060	0.820 ± 0.001	0.510 ± 0.005	0.983 ± 0.002
GPT-5.4-nano	4	**0.374 ± 0.017**	0.766 ± 0.007	0.583 ± 0.007	0.815 ± 0.021	0.456 ± 0.006	0.783 ± 0.006
GPT-5.4-nano	8	0.350 ± 0.018	0.554 ± 0.014	0.610 ± 0.062	0.866 ± 0.003	0.459 ± 0.014	0.912 ± 0.012
GPT-5.4-nano	16	0.323 ± 0.010	0.492 ± 0.002	0.600 ± 0.057	0.883 ± 0.002	0.389 ± 0.013	0.855 ± 0.012
Qwen-Inst	4	0.344 ± 0.001	0.687 ± 0.001	0.644 ± 0.000	0.943 ± 0.002	0.623 ± 0.000	0.926 ± 0.000
Qwen-Inst	8	0.338 ± 0.000	0.487 ± 0.002	0.644 ± 0.000	0.846 ± 0.000	0.627 ± 0.000	0.983 ± 0.000
Qwen-Inst	16	0.341 ± 0.001	0.441 ± 0.003	0.644 ± 0.001	0.896 ± 0.003	0.627 ± 0.000	0.959 ± 0.000
Qwen-Thinking	4	0.350 ± 0.009	0.748 ± 0.004	0.613 ± 0.001	0.968 ± 0.006	0.547 ± 0.007	0.942 ± 0.013
Qwen-Thinking	8	0.342 ± 0.007	0.530 ± 0.007	0.555 ± 0.038	0.833 ± 0.058	0.556 ± 0.006	0.839 ± 0.002
Qwen-Thinking	16	0.337 ± 0.001	0.495 ± 0.004	0.557 ± 0.044	0.850 ± 0.067	0.559 ± 0.007	0.814 ± 0.003
Qwen-Thinking-Budget	4	0.353 ± 0.008	0.733 ± 0.001	0.613 ± 0.005	0.966 ± 0.000	0.546 ± 0.017	0.952 ± 0.005
Qwen-Thinking-Budget	8	0.340 ± 0.004	0.533 ± 0.004	0.613 ± 0.001	0.919 ± 0.009	0.551 ± 0.003	0.972 ± 0.004
Qwen-Thinking-Budget	16	0.332 ± 0.003	0.497 ± 0.012	0.614 ± 0.002	0.945 ± 0.003	0.551 ± 0.003	0.941 ± 0.005

Table 16: Per-turn state-tracking and retrieval tasks: modality split performance. Tasks are partitioned based on the presence of images in the dialogue history.

Model	Ctx	DST-T	DST-M	ImgRet-T	ImgRet-M	TxtRet-T	TxtRet-M
Gemini-3-Flash	4	0.905 ± 0.011	0.231 ± 0.079	0.725 ± 0.011	0.832 ± 0.012	0.704 ± 0.006	0.600 ± 0.003
Gemini-3-Flash	8	0.914 ± 0.008	0.437 ± 0.113	0.868 ± 0.012	0.916 ± 0.007	0.731 ± 0.009	0.531 ± 0.008
Gemini-3-Flash	16	0.970 ± 0.009	0.674 ± 0.037	0.880 ± 0.012	0.928 ± 0.003	0.729 ± 0.015	0.611 ± 0.008
Gemini-3.1-Pro	4	0.591 ± 0.004	0.071 ± 0.008	0.804 ± 0.006	0.855 ± 0.015	0.705 ± 0.002	0.563 ± 0.010
Gemini-3.1-Pro	8	0.945 ± 0.002	0.236 ± 0.006	**0.912 ± 0.007**	0.924 ± 0.009	0.725 ± 0.007	0.573 ± 0.007
Gemini-3.1-Pro	16	0.968 ± 0.003	0.677 ± 0.007	0.872 ± 0.059	**0.936 ± 0.003**	0.697 ± 0.022	**0.640 ± 0.003**
Claude-4.5-Sonnet	4	0.733 ± 0.159	0.164 ± 0.087	0.769 ± 0.009	0.721 ± 0.004	0.695 ± 0.002	0.572 ± 0.006
Claude-4.5-Sonnet	8	0.893 ± 0.008	0.331 ± 0.118	0.845 ± 0.007	0.796 ± 0.014	0.704 ± 0.009	0.585 ± 0.002
Claude-4.5-Sonnet	16	0.935 ± 0.017	0.634 ± 0.017	0.859 ± 0.014	0.851 ± 0.016	0.697 ± 0.007	0.633 ± 0.008
GPT-5.4	4	0.763 ± 0.164	0.191 ± 0.109	0.745 ± 0.007	0.792 ± 0.009	0.725 ± 0.002	0.567 ± 0.004
GPT-5.4	8	0.945 ± 0.004	0.407 ± 0.147	0.869 ± 0.004	0.881 ± 0.002	**0.737 ± 0.005**	0.579 ± 0.008
GPT-5.4	16	**0.977 ± 0.008**	**0.775 ± 0.010**	0.871 ± 0.005	0.911 ± 0.012	**0.737 ± 0.004**	0.616 ± 0.007
GPT-5.4-nano	4	0.620 ± 0.180	0.135 ± 0.066	0.373 ± 0.010	0.405 ± 0.028	0.680 ± 0.003	0.521 ± 0.008
GPT-5.4-nano	8	0.697 ± 0.018	0.255 ± 0.089	0.537 ± 0.018	0.429 ± 0.022	0.680 ± 0.009	0.552 ± 0.015
GPT-5.4-nano	16	0.718 ± 0.040	0.403 ± 0.035	0.537 ± 0.017	0.411 ± 0.025	0.671 ± 0.005	0.589 ± 0.010
Qwen-Inst	4	0.741 ± 0.123	0.193 ± 0.069	0.713 ± 0.002	0.809 ± 0.002	0.691 ± 0.002	0.553 ± 0.002
Qwen-Inst	8	0.650 ± 0.124	0.336 ± 0.098	0.825 ± 0.002	0.847 ± 0.005	0.716 ± 0.006	0.555 ± 0.004
Qwen-Inst	16	0.684 ± 0.130	0.426 ± 0.048	0.829 ± 0.002	0.864 ± 0.000	0.715 ± 0.002	0.609 ± 0.002
Qwen-Thinking	4	0.319 ± 0.019	0.024 ± 0.003	0.705 ± 0.005	0.733 ± 0.013	0.693 ± 0.005	0.517 ± 0.007
Qwen-Thinking	8	0.461 ± 0.012	0.104 ± 0.009	0.848 ± 0.014	0.827 ± 0.009	0.708 ± 0.012	0.545 ± 0.002
Qwen-Thinking	16	0.456 ± 0.012	0.340 ± 0.009	0.867 ± 0.005	0.883 ± 0.013	0.708 ± 0.009	0.593 ± 0.008
Qwen-Thinking-Budget	4	0.435 ± 0.005	0.032 ± 0.003	0.745 ± 0.004	0.757 ± 0.008	0.701 ± 0.010	0.547 ± 0.008
Qwen-Thinking-Budget	8	0.669 ± 0.018	0.151 ± 0.007	0.861 ± 0.010	0.840 ± 0.007	0.717 ± 0.002	0.555 ± 0.012
Qwen-Thinking-Budget	16	0.696 ± 0.018	0.417 ± 0.008	0.861 ± 0.008	0.885 ± 0.005	0.716 ± 0.011	0.632 ± 0.014

Table 17: Cross-turn extraction tasks: modality split performance. Tasks are partitioned based on whether the target cluster involves image regions. A '*' superscript marks cells whose eval set contains fewer than 250 samples (smaller pool, see Appendix).

Model	Ctx	Span-T	Span-M	Coref-T	Coref-M	Rel-T	Rel-M
gemini-3-flash	4	0.556 ± 0.007	0.841 ± 0.015	0.419 ± 0.002	0.274 ± 0.028*	0.816 ± 0.005	0.333 ± 0.000*
gemini-3-flash	8	0.632 ± 0.011	0.828 ± 0.015	0.226 ± 0.007	0.254 ± 0.012	0.472 ± 0.007	0.553 ± 0.020
gemini-3-flash	16	**0.734 ± 0.010**	0.855 ± 0.003	0.231 ± 0.001	0.075 ± 0.005	0.272 ± 0.006	0.322 ± 0.010
gemini-3.1-pro	4	0.330 ± 0.006	0.847 ± 0.027	0.483 ± 0.013	**0.689 ± 0.031***	**0.949 ± 0.001**	0.333 ± 0.000*
gemini-3.1-pro	8	0.408 ± 0.004	0.872 ± 0.006	0.230 ± 0.014	0.588 ± 0.010	0.512 ± 0.006	**0.564 ± 0.012**
gemini-3.1-pro	16	0.489 ± 0.006	**0.885 ± 0.003**	0.063 ± 0.005	0.261 ± 0.009	0.193 ± 0.003	0.374 ± 0.009
claude-4.5-sonnet-20250929	4	0.353 ± 0.008	0.602 ± 0.004	0.416 ± 0.009	0.222 ± 0.031*	0.921 ± 0.003	0.222 ± 0.000*
claude-4.5-sonnet-20250929	8	0.456 ± 0.006	0.588 ± 0.007	0.184 ± 0.010	0.126 ± 0.003	0.405 ± 0.011	0.271 ± 0.014
claude-4.5-sonnet-20250929	16	0.585 ± 0.012	0.589 ± 0.003	0.196 ± 0.002	0.074 ± 0.003	0.215 ± 0.005	0.214 ± 0.004
gpt-5.4	4	0.311 ± 0.009	0.623 ± 0.009	**0.506 ± 0.005**	0.496 ± 0.028*	0.924 ± 0.005	0.222 ± 0.157*
gpt-5.4	8	0.359 ± 0.007	0.617 ± 0.004	0.167 ± 0.001	0.396 ± 0.005	0.391 ± 0.004	0.135 ± 0.007
gpt-5.4	16	0.419 ± 0.008	0.607 ± 0.009	0.109 ± 0.001	0.108 ± 0.008	0.188 ± 0.001	0.240 ± 0.002
gpt-5.4-nano	4	0.329 ± 0.001	0.444 ± 0.009	0.305 ± 0.007	0.111 ± 0.031*	0.919 ± 0.005	0.148 ± 0.105*
gpt-5.4-nano	8	0.445 ± 0.008	0.448 ± 0.006	0.086 ± 0.011	0.009 ± 0.005	0.377 ± 0.008	0.186 ± 0.008
gpt-5.4-nano	16	0.579 ± 0.006	0.423 ± 0.002	0.062 ± 0.007	0.004 ± 0.002	0.101 ± 0.006	0.091 ± 0.003
qwen-3-vl-32b-instruct	4	0.335 ± 0.002	0.516 ± 0.003	0.493 ± 0.002	0.200 ± 0.000*	0.891 ± 0.002	0.222 ± 0.000*
qwen-3-vl-32b-instruct	8	0.489 ± 0.003	0.545 ± 0.001	0.143 ± 0.000	0.069 ± 0.001	0.325 ± 0.003	0.247 ± 0.001
qwen-3-vl-32b-instruct	16	0.621 ± 0.007	0.522 ± 0.001	0.064 ± 0.002	0.002 ± 0.000	0.174 ± 0.001	0.212 ± 0.001
qwen-3-vl-32b-thinking	4	0.037 ± 0.004	0.137 ± 0.007	0.357 ± 0.023	0.267 ± 0.054*	0.734 ± 0.009	0.185 ± 0.139*
qwen-3-vl-32b-thinking	8	0.017 ± 0.007	0.117 ± 0.016	0.159 ± 0.009	0.196 ± 0.013	0.287 ± 0.007	0.191 ± 0.004
qwen-3-vl-32b-thinking	16	0.014 ± 0.003	0.111 ± 0.011	0.012 ± 0.001	0.006 ± 0.002	0.017 ± 0.006	0.075 ± 0.011
qwen-3-vl-32b-thinking-budget	4	0.341 ± 0.006	0.531 ± 0.004	0.326 ± 0.014	0.311 ± 0.063*	0.692 ± 0.017	0.222 ± 0.157*
qwen-3-vl-32b-thinking-budget	8	0.465 ± 0.015	0.537 ± 0.007	0.110 ± 0.005	0.269 ± 0.015	0.327 ± 0.010	0.377 ± 0.011
qwen-3-vl-32b-thinking-budget	16	0.615 ± 0.004	0.530 ± 0.002	0.018 ± 0.006	0.039 ± 0.007	0.228 ± 0.005	0.284 ± 0.008

D.3 Sparsity-robust coreference metrics

Exact-cluster F1 (our primary coreference metric) is sensitive to sparsity: when both the gold and the prediction contain no cluster it trivially scores 1.0, and this empty-vs-empty case is common at short context (e.g., 92.4% of text coreference instances have an empty gold at $k{=}4$, since few coreference chains fit inside a 4-turn window). Table 18 therefore reports two sparsity-robust metrics, B^3 and LEA (Moosavi and Strube, 2016), alongside exact-cluster F1 on the non-trivial subset (instances whose gold contains at least one cluster), for the same (model, context) cells as the Coref column of Table 14. The context-driven decline persists on this subset, indicating that it reflects genuinely more and larger coreference chains at longer context rather than only the empty-gold trivial cases.

To quantify how often exact-cluster F1 includes trivial empty–empty cases, Table 19 reports empty-gold, empty-prediction, and empty–empty rates. Across both modalities, three context lengths, and three reruns, 24.9% of evaluated cases have no gold cluster. This gold-empty rate is shared across models because they use the same samples; empty-prediction and empty–empty rates depend on model outputs. The rates describe the empty cases included in the primary scores and complement the non-empty-gold evaluation in Table 18.

D.4 Relation F1 with and without PART_OF

To assess whether the sparse PART_OF predicate affects the reported relation F1, we compare scores with and without it. PART_OF is the least frequent annotated predicate, with 737 instances (0.51% of all relations; Table 2).

Excluding PART_OF changes relation F1 by at most 0.0158 across the evaluated model–context combinations. For Table 20, we remove PART_OF triples from both predictions and gold and re-score the same model outputs on all evaluation samples. We retain the scoring rule that assigns $\text{F1} = 1$ when both sets of triples are empty. Sample-level F1 scores are averaged over both modalities and three reruns, matching the aggregation in Table 14. This comparison measures the effect of excluding PART_OF while keeping the evaluated samples fixed.

Table 18: Sparsity-robust coreference metrics on the *non-trivial* subset (instances whose gold contains at least one cluster; the empty-gold instances that exact-cluster F1 trivially credits are excluded). Companion to the **Coref** column of Table 14: same model/context cells, restricted to instances that actually require coreference. N is the number of such instances (text coreference gold is context-gated, so N grows with context). We report exact-cluster F1, B^3, and LEA.

Model	Ctx	N	Exact-F1	B^3-F1	LEA-F1
Gemini-3-Flash	4	102	0.419	0.808	0.577
	8	1248	0.269	0.704	0.494
	16	1500	0.153	0.615	0.341
Gemini-3.1-Pro	4	102	0.549	0.850	0.614
	8	1248	0.418	0.791	0.577
	16	1500	0.162	0.686	0.368
Claude-4.5-Sonnet	4	102	0.312	0.793	0.511
	8	1248	0.171	0.717	0.449
	16	1500	0.135	0.649	0.316
GPT-5.4	4	102	0.441	0.839	0.569
	8	1248	0.297	0.751	0.500
	16	1500	0.108	0.651	0.327
GPT-5.4-nano	4	102	0.193	0.729	0.319
	8	1248	0.052	0.627	0.265
	16	1500	0.033	0.584	0.184
Qwen-Inst	4	102	0.255	0.766	0.503
	8	1248	0.080	0.609	0.386
	16	1500	0.033	0.524	0.229
Qwen-Thinking	4	102	0.232	0.741	0.366
	8	1248	0.138	0.633	0.292
	16	1500	0.008	0.545	0.149
Qwen-Thinking-Budget	4	102	0.289	0.770	0.458
	8	1248	0.209	0.694	0.454
	16	1500	0.029	0.570	0.281

Model	Empty gold (%)	Empty prediction (%)	Both empty (%)
Gemini-3-Flash	24.90	8.62	8.12
Gemini-3.1-Pro	24.90	15.65	11.28
Claude-4.5-Sonnet	24.90	9.06	8.14
GPT-5.4	24.90	11.94	10.78
GPT-5.4-nano	24.90	5.90	5.82
Qwen-Inst	24.90	11.91	10.88
Qwen-Thinking	24.90	39.87	9.54
Qwen-Thinking-Budget	24.90	6.85	6.67

Table 19: Coreference empty-case rates across both settings, all three context lengths, and three reruns. Each model produces 3,795 predictions from 1,265 evaluation samples, counting each run separately. The columns report the percentages with no gold cluster, no predicted cluster, and neither, respectively. Cases with neither receive exact-cluster F1 $= 1$; cases with no gold cluster are excluded from Table 18.

	$k = 4$		$k = 8$		$k = 16$	
Model	All	No PART_OF	All	No PART_OF	All	No PART_OF
Gemini-3-Flash	0.8098	0.8125	0.5124	0.5174	0.2968	0.3025
Gemini-3.1-Pro	0.9415	0.9415	0.5383	0.5424	0.2833	0.2874
Claude-4.5-Sonnet	0.9127	0.9131	0.3376	0.3466	0.2146	0.2209
GPT-5.4	0.9160	0.9160	0.2628	0.2646	0.2139	0.2173
GPT-5.4-nano	0.9097	0.9097	0.2815	0.2824	0.0960	0.0966
Qwen-Inst	0.8830	0.8833	0.2857	0.2987	0.1932	0.2028
Qwen-Thinking	0.7273	0.7273	0.2413	0.2462	0.0446	0.0461
Qwen-Thinking-Budget	0.6859	0.6908	0.3519	0.3677	0.2561	0.2658

Table 20: Relation F1 with all predicates and after removing PART_OF from predictions and gold. At $k = 4, 8, 16$, the evaluation uses 253, 500, and 500 samples across both settings, respectively, with three reruns per sample. Both columns for each k average sample-level F1 over the same samples and runs. Samples with empty predictions and gold are retained and receive F1 $= 1$.

D.5 Controlled image ablation

To measure image dependence on the *same* instances (rather than the image-presence partition of the modality-split tables), we run a controlled ablation on identical samples at k=16: **image-ON** (raw history images), **image-OFF** (history image turns removed), and **image-as-text** (each history image turn replaced by its K-Fashion attribute tags). Table 21 reports the Condition-A (dialogue) tasks and Table 22 the Condition-B (extraction) tasks.

D.6 Supervised baseline training and evaluation

Training setup. To reproduce the supervised baselines in Section 6.5, we use the following training setup. Starting from Qwen3-VL-2B-Instruct, we train a separate LoRA (Hu et al., 2022) adapter for sentiment, intent, and strategy classification. Each task uses 5,000 training and 500 validation examples at k=16, randomly sampled from annotated turns, with equal numbers of text-only and multimodal examples; source dialogues are disjoint across training, validation, and evaluation. We apply LoRA to the language model's attention and feed-forward projections (r=16, α=32, dropout 0.05), keeping the original model weights and vision encoder fixed. Training uses AdamW with learning rate 10^{-4}, batch size 32, and three epochs with seed 42; only the answer JSON and end-of-sequence token contribute to the loss.

Evaluation protocol. We isolate the effect of supervised adaptation by comparing each adapter with the same unadapted checkpoint on identical evaluation samples. Both use the existing task prompts and the fixed k=16 evaluation pools (N=250 per task and setting), with greedy decoding and a 128-token output limit. Images are resized to at most 256 visual tokens each in both conditions. We select the checkpoint with the highest Macro-F1 on the 500 validation examples and evaluate it once on each test pool. Table 4 reports the resulting comparison.

E Output Failure Mode Breakdown

Aggregate failure rates per model are summarized in Table 23; per-task breakdowns for empty predictions, truncated outputs, and JSON-format failures follow.

F Comparison with Related Resources

Table 27 consolidates the axes of comparison referenced in Section 2: language (Korean), modality (multimodal), interaction format (multi-turn dialogue), the presence of coordinated fine-grained NLU annotations on the same data, and explicit cross-modal relation annotation. KoMMERCE is the only resource we surveyed that satisfies all five axes. We additionally include three customer-service dialogue corpora (JDDC, JDDC 2.1, CMCC) to cover scale-relevant prior work and report average utterances per dialogue (where applicable) as a single comparable scale axis.

G Collection-stage Guidelines and Quality Control Detail

This appendix collects the per-stage annotation guidelines that govern the three-stage collection pipeline summarized in Section 4, the full Data Quality Control workflow, and the dialogue-generation interface.

Utterance guideline (Stage 1). To balance goal orientation with natural variation, our guideline enforces: (i) open-ended exchanges that preserve the initial objective and the user's preferences across turns; (ii) system utterances restricted to the assigned strategy, with no unsolicited opinions; (iii) at least one fashion knowledge passage drawn from retrieved documents and at least one relevant image per session.

State guideline (Stage 2-1). Each fashion attribute is formatted as `polarity` + (`item` : `attribute-category` : `attribute-value`). For example, `top:category:blouse` encodes a requested blouse. (i) Polarity is positive by default; negative attributes prepend `!` for exclusion. (ii) Unspecified parts may be omitted, e.g., `(top: :)`. (iii) Item and attribute categories follow the K-Fashion taxonomy (AI-Hub, 2020) (1,992 total attributes).

Span guideline (Stage 2-2). Spans cover any semantically meaningful unit in the fashion domain (not limited to nouns / noun phrases): adjectives, attributive modifiers, adverbs of degree, compound forms, pronouns, and wh-words (e.g., "dress", "linen A-line dress", "it", "which", "slightly cropped", "what", "how"). For item mentions, the annotator first identifies the head (e.g., "dress") and then includes the modifiers needed to

Table 21: Three-condition image ablation on Condition-A (dialogue) tasks at context length 16. **image-ON** is the baseline with history image turns present; **image-OFF** removes them; **image-as-text** replaces each history image turn with its K-Fashion attribute tags as text. For image retrieval, only context image is dropped/textualized; candidate images are kept under all three conditions. Each cell is mean ± std over 3 reruns. Deltas are reported in `ablation_image_compare.csv`.

Model	Setting	Sent-M	Intent-M	Strat-M	DST-M	ImgRet-M	TxtRet-M
gemini-3-flash	image-ON	0.501 ± 0.009	0.967 ± 0.004	0.986 ± 0.001	0.643 ± 0.016	0.928 ± 0.003	0.611 ± 0.008
	image-OFF	0.527 ± 0.006	0.964 ± 0.002	0.985 ± 0.001	0.692 ± 0.014	0.895 ± 0.004	0.604 ± 0.011
	image-as-text	0.530 ± 0.012	0.971 ± 0.005	0.984 ± 0.002	0.648 ± 0.009	0.909 ± 0.007	0.629 ± 0.010
gemini-3.1-pro	image-ON	0.625 ± 0.004	0.953 ± 0.001	0.985 ± 0.001	0.723 ± 0.014	0.936 ± 0.003	0.640 ± 0.003
	image-OFF	0.617 ± 0.009	0.969 ± 0.003	0.939 ± 0.067	0.753 ± 0.015	0.919 ± 0.008	0.631 ± 0.005
	image-as-text	0.620 ± 0.004	0.964 ± 0.002	0.987 ± 0.000	0.709 ± 0.004	0.927 ± 0.004	0.625 ± 0.007
qwen-3-vl-32b-instruct	image-ON	0.441 ± 0.003	0.896 ± 0.003	0.959 ± 0.000	0.477 ± 0.002	0.864 ± 0.000	0.609 ± 0.002
	image-OFF	0.458 ± 0.001	0.913 ± 0.000	0.955 ± 0.000	0.484 ± 0.003	0.849 ± 0.002	0.580 ± 0.003
	image-as-text	0.464 ± 0.000	0.925 ± 0.003	0.951 ± 0.002	0.407 ± 0.002	0.883 ± 0.004	0.604 ± 0.003
qwen-3-vl-32b-thinking	image-ON	0.496 ± 0.004	0.804 ± 0.008	0.814 ± 0.003	0.353 ± 0.022	0.884 ± 0.014	0.589 ± 0.004
	image-OFF	0.504 ± 0.008	0.947 ± 0.006	0.809 ± 0.005	0.345 ± 0.007	0.857 ± 0.013	0.596 ± 0.006
	image-as-text	0.501 ± 0.004	0.906 ± 0.066	0.820 ± 0.003	0.344 ± 0.007	0.873 ± 0.008	0.599 ± 0.010
qwen-3-vl-32b-thinking-budget	image-ON	0.497 ± 0.012	0.945 ± 0.003	0.941 ± 0.005	0.421 ± 0.008	0.885 ± 0.005	0.632 ± 0.014
	image-OFF	0.495 ± 0.014	0.942 ± 0.004	0.949 ± 0.008	0.472 ± 0.011	0.889 ± 0.008	0.624 ± 0.013
	image-as-text	0.508 ± 0.010	0.953 ± 0.001	0.960 ± 0.005	0.432 ± 0.009	0.895 ± 0.008	0.641 ± 0.008

uniquely specify the referent (e.g., "beige linen dress"). When mentions overlap, the most informative minimally sufficient span is chosen (e.g., "summer linen dress" over "linen dress"). Annotators favor spans aligned with the K-Fashion taxonomy across category, attribute, and style.

Crowd worker training. To instruct crowd workers in using the collection platform and annotating dialogues, we implemented a structured training program. We drafted guidelines and iteratively revised them three times in consultation with the professional data-collection vendor. Following a pilot with 20 workers, we validated the guidelines through trial tasks and updated them to address observed failures and ambiguities. We then scaled to 100 workers, delivered training over three months, and assigned roles based on demonstrated mastery. High-performing workers were promoted to quality reviewers who reviewed collected data, while others served as annotators or dialogue participants.

Data source control. We imposed strict control over data sources to address licensing constraints and to enhance the reliability of grounded texts and images. Unrestricted image use in the fashion domain can raise copyright and portrait-rights concerns; to mitigate these risks, we restrict all images to the K-Fashion dataset, whose contents are license-cleared, and remove explicit personal identifiers through cropping and blurring. For textual grounding, we limited knowledge sources to publicly accessible Wikipedia (Wikipedia contributors, 2024) and NamuWiki (NamuWiki contributors, 2024) (a widely used Korean community wiki, noting that NamuWiki is community-edited rather than formally verified).

Quality review. At each stage, dialogue samples advanced only after passing two sequential quality reviews, termed Level 1 and Level 2. In Level 1, reviewers assessed whether the dialogues were fit for the task requirements and the stage-specific guidelines. When issues were identified, reviewers either corrected the dialogue directly or returned the sample to the original speakers or annotator with revision requests. Level 2 served as a second pass to double-check residual errors and finalize the sample; once Level 2 passed, the dialogue was locked to prevent further edits and to avoid downstream inconsistencies when subsequent annotation stages were added.

Dialogue generation interface. Figure 16 shows the role-based UI used by dialogue speakers. The user side (left) presents the topic, dialogue history, and a free-text input box; the system side (right) provides the same with built-in retrieval tools for text knowledge and K-Fashion images, plus widgets for tagging retrieved evidence to the current utterance.

H Per-task Dataset Statistics

This appendix collects the per-task distributions and figures referenced from Section 5.

Table 22: Image-as-text ablation on Condition-B (extraction) tasks at context length 16. Image mentions (bbox) in the pool are reverse-mapped to a K-Fashion part and replaced with that part's attribute tags as text; history image turns are also rendered as image-level tags. No raw image is provided. **image-ON** is the baseline (raw images shown). Δ = as-text − ON; a near-zero Δ means the model solves cross-modal binding via attribute-text matching, a negative Δ indicates the model relies on raw visual signal beyond what tags capture. span_detection is excluded (bbox annotations come from a different source than `bbox_info`). Each cell is mean ± std over 3 reruns.

Model	Setting	Coref-M	Rel-M
gemini-3-flash	image-ON	0.075 ± 0.005	0.322 ± 0.010
	image-as-text	0.080 ± 0.004	0.334 ± 0.011
	Δ	+0.005	+0.012
gemini-3.1-pro	image-ON	0.260 ± 0.009	0.374 ± 0.009
	image-as-text	0.197 ± 0.005	0.382 ± 0.003
	Δ	-0.064	+0.008
qwen-3-vl-32b-instruct	image-ON	0.002 ± 0.000	0.212 ± 0.002
	image-as-text	0.001 ± 0.000	0.246 ± 0.001
	Δ	-0.001	+0.034
qwen-3-vl-32b-thinking	image-ON	0.007 ± 0.003	0.072 ± 0.009
	image-as-text	0.017 ± 0.004	0.050 ± 0.002
	Δ	+0.010	-0.022
qwen-3-vl-32b-thinking-budget	image-ON	0.039 ± 0.007	0.284 ± 0.008
	image-as-text	0.049 ± 0.002	0.280 ± 0.002
	Δ	+0.010	-0.003

Table 23: Overall output-failure rates per model, aggregated over all 18 task variants × 3 context lengths × 3 reruns. Empty% is over all predictions; truncation% and JSON-fail% are over samples with a recorded raw-response length.

Model	#pred	Empty%	Trunc%	JSON-fail%
Gem-Flash	39,054	3.6	0.0	2.7
Gem-Pro	39,054	7.0	0.0	3.0
Claude-4.5-Sonnet	39,054	4.3	0.0	4.3
GPT-5.4	39,054	5.1	0.0	1.8
GPT-5.4-nano	39,054	5.7	0.0	4.2
Qwen-Inst	39,054	4.3	0.0	2.0
Qwen-Think	39,054	24.5	16.2	17.9
Qwen-Think-Bgt	39,054	2.8	0.0	1.8

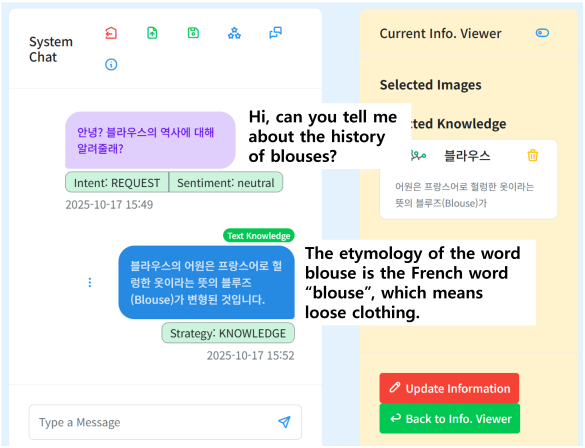


Figure 16: Dialogue generation interface. **Left:** a user (purple) and a system (blue) perform a role-play to pursue the initial objective about "blouse." **Right:** the system may attach related images or text knowledge.

Fashion attribute hierarchy. The images are organized along a hierarchical attribute taxonomy (Figure 17); each percentage reflects the share of an attribute over total annotations. The depth of the hierarchy supports fine-grained image retrieval evaluation: candidates must be matched on detailed sub-attributes (e.g., sleeve style, neckline) rather than coarse category alone.

Intent and sentiment distributions. Figure 18 visualizes the intent (a) and sentiment (b) distributions. Intent is dominated by REFINED-REQUEST (28.3%) and CHAT (27.7%), with EVALUATION (14.5%), KNOWLEDGE (11.5%), REQUEST (10.3%), and DETAIL (7.8%) covering the remainder. Sentiment reflects the structure of shopping conversations: most informational exchanges are neutral, while positive and negative reactions cluster around recommendation moments.

Strategy distribution. CLARIFICATION accounts for the largest share of system strategies (Figure 19), reflecting that user requests typically arrive in incomplete form and the system must specify preferences before recommending. DETAIL is

Table 24: Empty-prediction rate (%) per task variant and model. A prediction is empty when parsing yields None/""/[]/{}. A '*' marks extraction tasks where an empty list may also be a valid 'no items' answer.

Task	Gem-Flash	Gem-Pro	Claude-4.5-Sonnet	GPT-5.4	GPT-5.4-nano	Qwen-Inst	Qwen-Think	Qwen-Think-Bgt
Sent-T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Sent-M	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
Intent-T	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0
Intent-M	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Strat-T	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0
Strat-M	0.0	0.4	0.0	0.0	0.0	0.0	0.3	0.0
DST-T	0.0	0.0	0.1	0.0	0.0	0.0	23.2	0.0
DST-M	0.0	0.0	0.0	0.0	0.0	0.0	17.0	0.0
ImgRet-T	0.1	1.4	0.0	0.8	0.8	0.0	3.2	0.0
ImgRet-M	0.2	0.4	0.0	0.1	0.1	0.0	3.8	0.0
TxtRet-T	0.0	0.8	0.0	0.0	0.0	0.0	6.1	0.0
TxtRet-M	0.0	0.3	0.0	0.0	0.0	0.0	7.3	0.0
Span-T*	5.1	22.7	7.0	13.4	4.3	7.3	97.0	1.4
Span-M*	0.1	1.0	0.0	0.0	0.0	0.0	79.3	0.0
Coref-T*	14.4	26.4	14.8	20.1	10.0	19.9	38.7	11.6
Coref-M*	0.3	0.1	0.7	0.0	0.0	0.3	41.6	0.0
Rel-T*	41.6	64.4	50.3	52.0	56.0	46.3	78.5	34.2
Rel-M*	1.5	4.2	3.4	2.5	41.4	1.0	63.5	1.7

Table 25: Token-truncation rate (%) per task variant and model. A response is truncated when its raw length exceeds 12,000 characters (≈ the 4096-token cap). '†' marks cells estimated from the sampled `examples` set (raw length not recorded in `all_samples`).

Task	Gem-Flash	Gem-Pro	Claude-4.5-Sonnet	GPT-5.4	GPT-5.4-nano	Qwen-Inst	Qwen-Think	Qwen-Think-Bgt
Sent-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Sent-M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Intent-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Intent-M	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0
Strat-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Strat-M	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0
DST-T	0.0	0.0	0.0	0.0	0.0	0.0	31.3	0.0
DST-M	0.0	0.0	0.0	0.0	0.0	0.0	20.2	0.0
ImgRet-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	1.1†	0.0
ImgRet-M	0.0	0.0	0.0	0.0	0.0	0.0	3.7	0.0
TxtRet-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	2.2†	0.0
TxtRet-M	0.0	0.0	0.0	0.0	0.0	0.0	4.6	0.0
Span-T	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0
Span-M	0.0	0.0	0.0	0.0	0.0	0.0	14.3	0.0
Coref-T	0.0	0.0	0.0	0.0	0.0	0.0	29.4	0.0
Coref-M	0.0	0.0	0.0	0.0	0.0	0.0	41.7	0.0
Rel-T	0.0	0.0	0.0	0.0	0.0	0.0	34.0	0.0
Rel-M	0.0	0.0	0.0	0.0	0.0	0.0	49.8	0.0

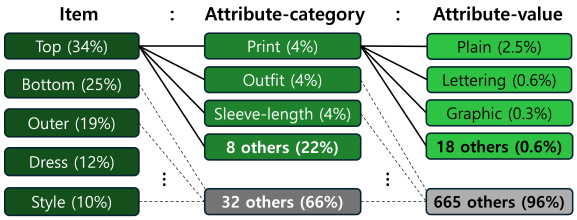


Figure 17: Hierarchical composition of fashion attributes assigned to images, which supports the image diversity. Percentages are computed over the total number of annotated attributes. "Others" aggregates the remaining categories.

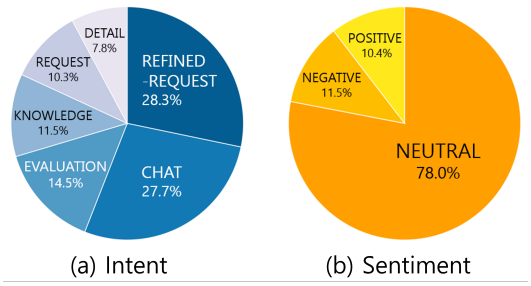


Figure 18: Distributions of user intent and sentiment.

the least frequent strategy but still includes more than 20,000 utterances, so no category suffers from data sparsity.

State tracking transitions. Query-state length per dialogue averages 5.04 attributes (maximum 14). State transitions partition into ADD (5.75 per dialogue on average), CHANGE (0.28), and RESET (0.01), indicating that users primarily refine re-

Table 26: JSON-format failure rate (%) per task variant and model. The model produced a non-trivial response (>50 chars, not truncated) that failed to parse as the expected JSON. '†' marks sampled cells.

Task	Gem-Flash	Gem-Pro	Claude-4.5-Sonnet	GPT-5.4	GPT-5.4-nano	Qwen-Inst	Qwen-Think	Qwen-Think-Bgt
Sent-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Sent-M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Intent-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Intent-M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Strat-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
Strat-M	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DST-T	0.0	0.0	0.1	0.0	0.0	0.0	0.5	0.0
DST-M	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
ImgRet-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	0.0†	0.0
ImgRet-M	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
TxtRet-T	0.0†	0.0†	0.0	0.0	0.0	0.0†	10.0†	0.0
TxtRet-M	0.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0
Span-T	0.0	0.0	7.0	11.0	0.0	5.2	95.4	0.4
Span-M	0.0	0.0	0.0	0.0	0.0	0.0	64.6	0.0
Coref-T	7.8	0.1	14.7	0.0	0.0	0.5	2.2	0.0
Coref-M	0.2	0.0	0.7	0.0	0.0	0.3	2.7	0.0
Rel-T	24.3	35.0	49.7	18.2	45.5	18.0	44.3	29.9
Rel-M	1.5	3.9	3.4	2.5	41.4	1.0	13.0	1.7

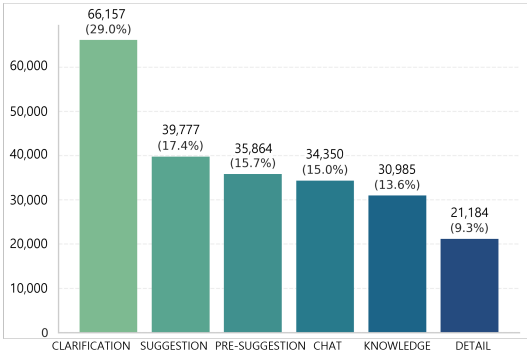


Figure 19: Distribution of strategies. All six strategies appear more than 20,000 times, indicating that no category suffers from data sparsity.

quirements incrementally; however, complex alternating transitions appear in over 20% of dialogues, supporting non-trivial DST evaluation under realistic conversational shifts.

Text retrieval frequency. The corpus contains an average of 2.03 text retrievals per dialogue, up to 7 in the longest case. The average number of utterances preceding a retrieval is 13.87, placing most retrievals in the early-to-middle stages of dialogues whose average length is 35.25 utterances.

Image retrieval attribute complexity. User retrieval utterances contain a median of 3 fashion attributes (75th percentile: 5; maximum: 10), while candidate images contain a median of 14 attributes (75th percentile: 26). Retrieval must therefore match attribute subsets rather than whole bags, making the task a non-trivial multi-constraint problem.

Span detection distributions. Figure 20 plots the distribution of text-span lengths and image-span area ratios. For text, 76.31% of spans (299,774) fall in the $[1, 10)$-character bin, reflecting the prevalence of deictic expressions ("this one"), object-level categories (e.g., dress), and fine-grained attributes (e.g., black, hem length) that dominate user-system references. Image spans show a flatter distribution of area ratios; excluding 481 images without bounding boxes, the remaining 32,454 images yield a median area ratio of 25.85%, consistent with full-body fashion photos where target items occupy approximately one quarter of the frame.

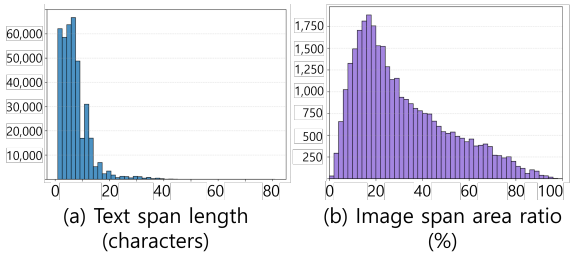


Figure 20: Distributions of text and image spans.

Coreference entity and mention-distance distributions. The corpus contains 181,497 entities after coreference resolution; 149,191 (82.20%) consist exclusively of text mentions, and 32,306 (17.80%) contain both text and image mentions. To quantify the breadth of textual context required, we measure the utterance distance between mentions within the same entity: 91.21% of coreferent pairs fall within 10 utterances, while multi-session dialogues produce rare long-range cases up to 59

Benchmark	Cite	Korean	Multi-modal	Multi-turn dialogue	Co-evaluated fine-NLU	Cross-modal relation	Avg utt./ dialogue
Korean LLM benchmarks (single-turn, text-only, MCQ)							
KMMLU	(Son et al., 2024)	✓					–
KMMLU-Redux/Pro	(Hong et al., 2025)	✓					–
KoCommonGEN v2	(Seo et al., 2024)	✓					–
KoBALT	(Shin et al., 2025)	✓					–
Multimodal dialogue benchmarks							
MMD	(Saha et al., 2018)		✓	✓			40
SIMMC 2.0	(Kottur et al., 2021)		✓	✓	partial		10.6
MMDialog	(Feng et al., 2023)		✓	✓			4.6
MUSE	(Wang et al., 2025b)		✓	✓			11.9
ECom-Bench	(Wang et al., 2025a)		✓	✓	end-to-end†		≤20‡
Customer-service dialogue corpora							
JDDC	(Chen et al., 2020)			✓			20
JDDC 2.1	(Zhao et al., 2022)		✓	✓			14.1
CMCC	(Huang et al., 2022)		✓	✓			22.31
Cross-modal NLU benchmarks (no dialogue)							
MNRE	(Zheng et al., 2021)		✓			✓	–
MORE	(He et al., 2023)		✓			✓	–
MRE-MI	(Huang et al., 2025)		✓			✓	–
KoMMERCE (ours)		✓	✓	✓	✓(9 tasks)	✓	**35.25**

Table 27: Axes of comparison between KoMMERCE and representative prior resources. "Co-evaluated fine-NLU" refers to having multiple NLU subtasks annotated on the *same* dialogues. SIMMC 2.0 is marked "partial" because it co-annotates a subset (state tracking, coreference, disambiguation) but not the broader nine-task suite we provide. ECom-Bench is marked "end-to-end†" because its primary metric is a single pass^3 outcome rather than per-NLU-competence scores. "Avg utt./dialogue" is reported by, or directly computed from, each original paper for scale comparison only (higher is not necessarily better; production-log corpora are short by design). "–" marks corpora where the number is not applicable. ECom-Bench‡ reports only an evaluation-time turn cap ($\leq$ 20 per task).

utterances apart.

Cross-modal relation mention-pair distance. Figure 21 reports the ECDF of head-to-tail utterance distance across all annotated cross-modal relations. The median distance is 3 and the 95th percentile is 17 utterances, confirming that relation extraction in KoMMERCE spans substantive dialogue contexts rather than collapsing into within-utterance triples.

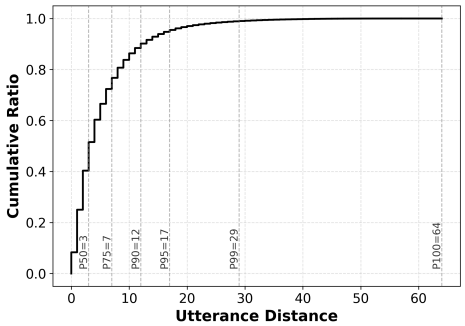


Figure 21: Cumulative ratio of relation mention-pair distances in dialogues. The distance denotes the absolute difference between the utterance indices of the head and tail mentions.

Image attribute field	Distinct values
Category	21
Color	21
Secondary color	21
Material	25
Print	21
Length	10
Sleeve length	6
Neckline	12
Collar	9
Fit	7
Detail	40
Style	23
Secondary style	23

Table 28: Image diversity within the fashion domain, quantified by distinct attribute values across 29,829 unique images. Values are deduplicated across garment categories; secondary color and style are counted separately.