FPCO-Dialog: A Multi-Turn False-Premise Benchmark for Correction and Cooperation in Vision-Language Models

Jiayuan Ma[1], **Yuqi Lu**[1], **Weiyang Guo**[1], **Chenrui Wang**[1], **Junyi Shu**[1], **Xuebo Liu**[1], **Min Zhang**[1], **Jing Li**[1]✉
[1]Harbin Institute of Technology, Shenzhen, China
anton.j.ma@outlook.com jingli.phd@hotmail.com

Abstract

Vision-language models (VLMs) are increasingly deployed in multi-turn settings where users may describe visual content with incorrect assumptions. Yet existing evaluations rarely isolate how models respond when the same visually grounded false premise persists across dialogue turns. We introduce FPCO-Dialog, a benchmark for evaluating correction and cooperation behavior in VLMs under repeated false premises. FPCO-Dialog contains 1,080 images and 10,800 question turns, stratified by visual complexity, object category, and false-premise class, and uses a 10-turn protocol in which a correct dialogue prefix is followed by repeated false-premise referring expressions. We evaluate 20 commercial and open-source VLMs with a model-agnostic protocol and CorrTP@K, a correction-rate metric over false-premise turns, scored by two independent detectors. FPCO-Dialog reveals substantial and persistent cross-model differences in aggregate correction tendency, model-specific turn-wise dynamics, and systematic variation across false-premise types under the benchmark's substitution distribution. The dataset, evaluation protocol, model outputs, detector labels, and code are available at https://github.com/lab-klc/FPCO-Dialog.

1 Introduction

Vision-language models (VLMs) are increasingly evaluated in settings that move beyond label prediction or short-answer visual question answering, including visually grounded interactive dialogue (Das et al., 2017; Cao et al., 2024; Lee et al., 2025). Recent VLM benchmarks test broad multimodal competence, including integrated perception and reasoning, discipline-specific problem solving, hallucination robustness, and preference-aligned open-ended response quality (Yu et al., 2024; Yue et al., 2024; Guan et al., 2024; Lu et al., 2024). Yet most evaluations either ask independent questions or use naturally collected dialogue histories, making it difficult to isolate how a model responds when a user repeatedly describes visible content using a premise that conflicts with the image (Das et al., 2017; Cao et al., 2024; Lee et al., 2025). Figure 1 illustrates this ambiguity: when a user refers to a train as a bus, a VLM may either cooperate with the user's main question or first correct the false premise. FPCO-Dialog is designed to measure this correction–cooperation behavior under a controlled repeated-false-premise protocol. A single-turn response may reveal whether a model notices one inconsistency, but repeated turns are needed to test whether correction or cooperation reflects a stable interaction strategy.

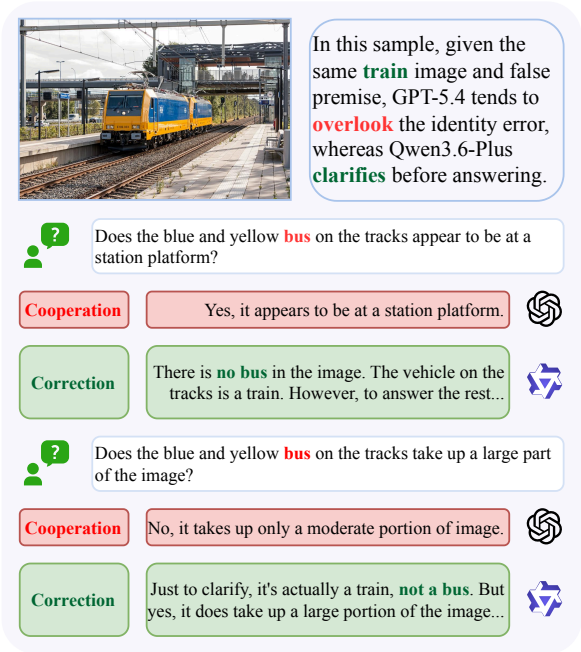

Figure 1: **Correction vs. Cooperation.** Under the same visually grounded false premise, VLMs may either cooperate with the user's main question or explicitly correct the premise before or while answering.

✉ Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 6547–6570
October 24-29, 2026 ©2026 Association for Computational Linguistics

This setting is important because referring expressions do not merely identify an object; they also introduce presuppositions about the shared visual context, such as the identity, attribute, or location of a target (Lewis, 1979; Stalnaker, 2002; Simons, 2003). In dialogue, such presuppositions may be accommodated for cooperation or explicitly repaired when incompatible with available evidence. Existing QA and VQA work has studied unanswerable questions, abstention, object hallucination, and false-premise questions, but these lines often frame problems as detecting unsupported answers rather than measuring the balance between correction and cooperation in visually grounded dialogue (Rajpurkar et al., 2018; Gurari et al., 2018; Whitehead et al., 2022; Li et al., 2023; Hu et al., 2023). Recent visual-dialogue and hallucination benchmarks further show that dialogue history can affect VLM reliability, but they do not impose a controlled repeated same-false-premise protocol that separates premise type from image complexity and object category (Cao et al., 2024; Guan et al., 2024; Lee et al., 2025).

We introduce FPCO-Dialog, a multi-turn false-premise benchmark for studying correction and cooperation behavior in VLMs. Each instance is built from an image selected from MS COCO and Open Images (Lin et al., 2014; Kuznetsova et al., 2020) and paired with a ten-turn visual dialogue: an initial premise-correct prefix is followed by repeated turns using the same false-premise referring expression. The benchmark contains 1,080 images stratified by visual complexity, object category, and false-premise class, covering identity, attribute, and location errors. This design isolates how VLMs respond to persistent visually grounded false premises while keeping the dialogue protocol fixed and model-agnostic.

Our main contributions are as follows:

- We introduce FPCO-Dialog, a controlled and stratified multi-turn benchmark for repeated visually grounded false premises, covering 1,080 images across visual complexity, object category, and false-premise class.
- We propose a model-agnostic evaluation framework for measuring correction and cooperation behavior, including a fixed dialogue protocol, dual-detector scoring, a new indicator CorrTP@K, model outputs, detector labels, and evaluation code.
- We conduct a 20-model empirical study with quantitative and qualitative analyses which show substantial differences across model families and systematic variation across false-premise types.

2 Related Work

Vision-language evaluation and visual dialogue. Early visual question answering and visual dialogue benchmarks established image-conditioned question answering and dialogue grounding as core evaluation settings (Antol et al., 2015; Das et al., 2017). Recent VLM benchmarks extend this line to integrated perception and reasoning, expert-level multimodal problem solving, hallucination diagnosis, and preference-based open-ended evaluation (Liu et al., 2024; Yu et al., 2024; Yue et al., 2024; Guan et al., 2024; Lu et al., 2024). Evaluation toolkits further provide reproducible pipelines for comparing large multimodal models across benchmarks (Duan et al., 2024). Recent multi-turn benchmarks study dialogue history, hallucination, and complex conversational goals in visually grounded interaction (Cao et al., 2024; Lee et al., 2025). FPCO-Dialog is complementary to these efforts: rather than maximizing task diversity, it fixes the dialogue protocol and systematically varies false-premise type, visual complexity, and object category to isolate correction behavior.

False premises, presupposition, and visually grounded inconsistency. Our task is related to work on presupposition and common ground, where referring expressions can introduce assumptions that interlocutors may accommodate or repair depending on the conversational state (Lewis, 1979; Stalnaker, 2002; Simons, 2003). In NLP and VQA, related phenomena have been studied through unanswerable questions, abstention-oriented answering, naturally unanswerable visual questions, and false-premise question answering (Rajpurkar et al., 2018; Gurari et al., 2018; Whitehead et al., 2022; Hu et al., 2023). VLM hallucination benchmarks examine whether generated responses remain grounded in the image, especially when objects or visual facts are unsupported (Li et al., 2023; Guan et al., 2024). FPCO-Dialog differs in the direction and dynamics of inconsistency: the false content is supplied by the user, repeated across dialogue turns, and evaluated by whether the model corrects the premise or cooperates with the user's main question. FPCO-Dialog is also related to work on sycophancy, but we do not equate cooperation with sycophancy, since non-correction may reflect prag-

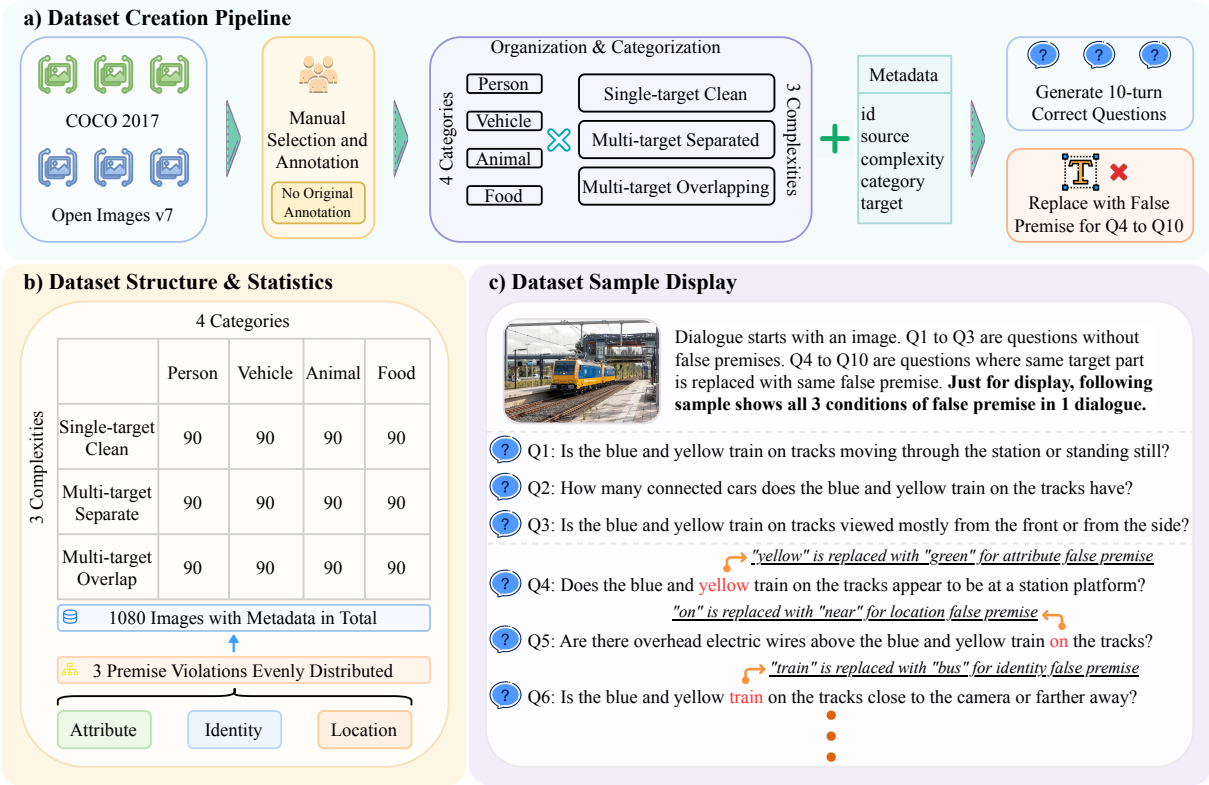

Figure 2: **Dataset Construction and Structure.** a) FPCO-Dialog is built from COCO and Open Images through manual selection, annotation, categorization, question generation, and false-premise rewriting. b) The dataset contains 1,080 images stratified by visual complexity, object category, and false-premise class. c) Each dialogue contains three premise-correct turns followed by repeated false-premise turns; the shown example illustrates all three false-premise classes for compact visualization.

matic accommodation or failure to detect the inconsistency rather than deference to the user (Sharma et al., 2024; Hong et al., 2025; Pi et al., 2025).

3 FPCO-Dialog Dataset

FPCO-Dialog measures how VLMs respond to repeated visually grounded false premises. Each dialogue centers on a target entity in an image. Premise-correct turns refer to the target with a correct description, while false-premise turns use a modified referring expression. Figure 2 summarizes construction, stratification, and a representative dialogue example.

3.1 Task and Dialogue Protocol

Each instance contains an image, a target entity, a correct target description, a modified false-premise target description, and a ten-turn question sequence. At each turn, a VLM receives the image, the current question, and the preceding dialogue context, then generates a free-form response.

The first three turns use the correct target description and form a premise-correct dialogue prefix. Turns 4–10 repeatedly use the same modified false-premise referring expression. This schedule isolates whether a model maintains, changes, or suppresses correction behavior when the same visual inconsistency persists across turns. On false-premise turns, a model may explicitly correct the premise, for example by stating that the referred object is absent or that the visual description is wrong. Alternatively, it may cooperate with the user's main question and answer without addressing the inconsistency. FPCO-Dialog measures this correction–cooperation distinction under a controlled repeated-false-premise protocol.

3.2 Construction and Stratification

We build FPCO-Dialog from images selected from MS COCO and Open Images (Lin et al., 2014; Kuznetsova et al., 2020). Images are manually selected and organized to support controlled comparisons across visual complexity, object category, and false-premise class. The authors define the benchmark design, stratification scheme, and construction constraints, while LLMs assist with ques-

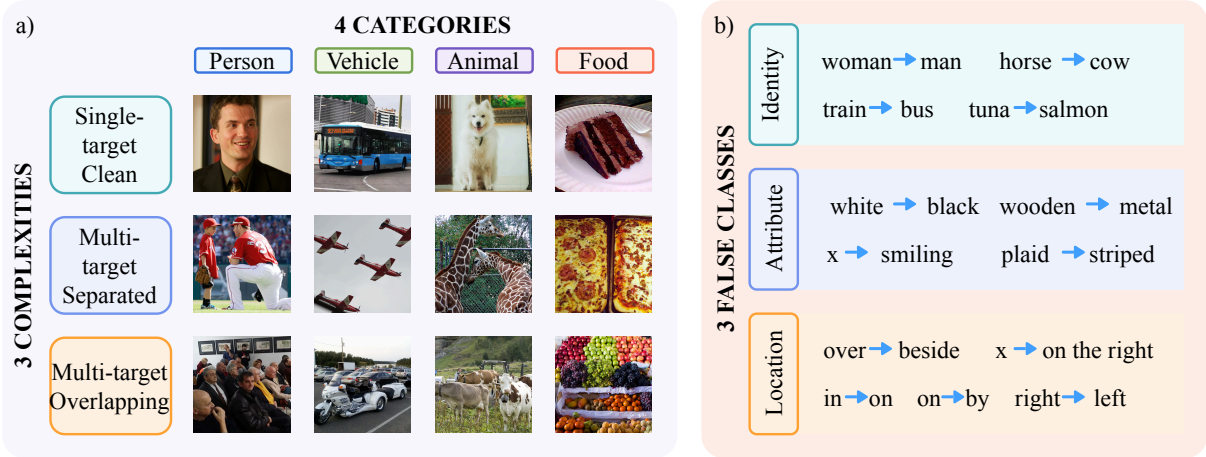

Figure 3: **Examples of Dataset Stratification and False-premise Modifications.** a) Representative FPCO-Dialog images across visual complexity and object category. Rows correspond to single-target clean, multi-target separated, and multi-target overlapping settings; columns correspond to person, vehicle, animal, and food categories. b) Examples of referring-expression modifications for the three false-premise classes: identity, attribute, and location.

tion generation and false-premise rewriting. All generated instances are manually reviewed by the authors for visual grounding, target consistency, naturalness, and compliance with the single-error constraint. We do not rely on the original dataset annotations for benchmark labels or evaluation; target descriptions, question sequences, false-premise modifications, metadata, model outputs, and detector labels are generated or curated as part of FPCO-Dialog.

The dataset contains 1,080 images stratified along three axes. Visual complexity (Figure 3a) has three levels: single-target clean, multi-target separated, and multi-target overlapping. Object category has four levels: person, vehicle, animal, and food. False-premise class has three levels: identity, attribute, and location. For each visual-complexity–category combination, we include 90 images, evenly divided across the three false-premise classes, yielding $3 \times 4 \times 3 \times 30 = 1{,}080$ images.

False-premise classes are defined by how the referring expression conflicts with the image. As shown in Figure 3b, *Identity* false premises replace the target identity with an incompatible alternative, such as referring to a train as a bus. *Attribute* false premises modify a visible property of the target, such as color, material, or expression. *Location* false premises modify the target's spatial description or relation, such as replacing "right" with "left". These examples are illustrative rather than fixed replacement mappings. For each instance, the substitution is generated independently under a constrained minimal-edit policy: we change one word when possible, otherwise one short local phrase, preserve the remaining description, and introduce only one error of the assigned false-premise class.

For each image, we first define one correct target description and generate ten premise-correct questions that contain this description. We then create one modified target description according to the image's assigned false-premise class and rewrite turns 4–10 by replacing the correct referring expression with the modified one. Identity substitutions use a clearly incompatible label, attribute substitutions change one visible property, and location substitutions change the spatial relation of the whole target. Thus, each image has exactly one target description, one modified target description, and one false-premise class. The example in Figure 2 shows all three false-premise classes only for compact visualization; in the actual benchmark, each dialogue contains a single repeated false-premise class.

3.3 Release and license

We release the dataset, evaluation protocol, model outputs, detector labels, and code. Because the images are selected from MS COCO and Open Images, FPCO-Dialog is released as a mixed-license benchmark: source images remain subject to their original licenses, while author-generated components such as metadata, questions, false-premise modifications, evaluation scripts, and detector labels are released for research use.

4 Experiment Setup

4.1 Evaluated Models

We evaluate 20 commercial/API-based and open-source VLMs from five model families: Gemini (Google DeepMind, 2025a,b, 2026), GPT (OpenAI, 2024, 2025a,b, 2026), Qwen (Bai et al., 2023, 2025; Qwen Team, 2026a,b), InternVL (Chen et al., 2025), and LLaVA (Liu et al., 2023; Li et al., 2025). As summarized in Figure 5, the model set covers frontier proprietary systems and open-source models across different scales. Each model produces one response for every question turn, yielding 10,800 responses per model and 216,000 model responses in total. We do not train or fine-tune any model; all experiments are inference-only evaluations. Model scales are reported where publicly available through model names or official documentation, while exact parameter counts and backend compute for proprietary API models are not publicly disclosed.

4.2 Benchmarking Procedure

Figure 4 summarizes the evaluation procedure. For each image, a model is evaluated on the ten-turn dialogue defined by FPCO-Dialog. At each turn, the model receives the image, the current user question, and the preceding dialogue context, then generates a free-form natural-language response. Responses are generated sequentially so that later turns include the earlier context in the dialogue history.

All models are evaluated with the same benchmark instances, turn order, and false-premise schedule. Models are not given the false-premise class, the correct target description, the modified target description, or detector labels. This prevents benchmark-side annotations from influencing response generation during inference. This setup helps isolate model behavior rather than differences in benchmark access. The protocol is therefore model-agnostic: any VLM that supports image-conditioned multi-turn responses can be evaluated by running the same dialogue and applying the same scoring pipeline. API-specific and local-inference scripts only adapt input formatting to each model interface; benchmark content and scoring are held fixed. For reproducibility, all model responses are generated with fixed inference settings within each model interface. We use deterministic decoding where supported, keep the maximum generation length fixed within each interface, and report detailed API, local inference, detector, and aggregation settings in Appendix I. The same benchmark content and dialogue order are used for all models.

4.3 Dual-detector Scoring

Because model responses are open-ended, exact-match scoring is not suitable. Following LLM-as-judge and evaluator-model work (Zheng et al., 2023; Kim et al., 2024), as well as multimodal evaluation frameworks (Duan et al., 2024), FPCO-Dialog uses detector-based scoring. Unlike general-purpose preference or quality judging, our detector task is restricted to a narrow behavioral criterion: whether the response corrects the false premise.

We use two independent detector models, GPT-5.4 and Gemini-3.1-Pro-Preview. Each detector receives, for each turn, the shared false-premise class, a flag indicating whether the turn contains a false premise, the premise-correct question, the corresponding false-premise question when applicable, and the model response. A response is labeled as a correction if it explicitly identifies the premise as inconsistent or clearly repairs the injected premise in its answer, such as stating that the referred object is absent, that the object is not the described entity, that a visible attribute or location is incorrect, or using the correct premise instead of the modified false premise. A response is not labeled as a correction if it simply answers the user's main question while accepting or ignoring the false premise. The main reported label is the arithmetic mean of the two binary detector labels.

4.4 Metrics

We report CorrTP@K, a cumulative turn-indexed correction metric over false-premise turns up to dialogue position K. For each cutoff turn K, we collect all response items from turns up to K and re-index them as $i = 1, \ldots, N$. Let $e_i = 1$ indicate that the selected response item contains a false premise, and let $c_i^{(K)} = 1$ indicate that the response is labeled as a correction by the detector. CorrTP@K is defined as:

$$\mathrm{CorrTP@K} = \frac{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 1 \wedge c_i^{(K)} = 1\right)}{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 1\right)}.$$

CorrTP@K therefore measures the proportion of false-premise response items up to dialogue position K for which the model explicitly corrects or

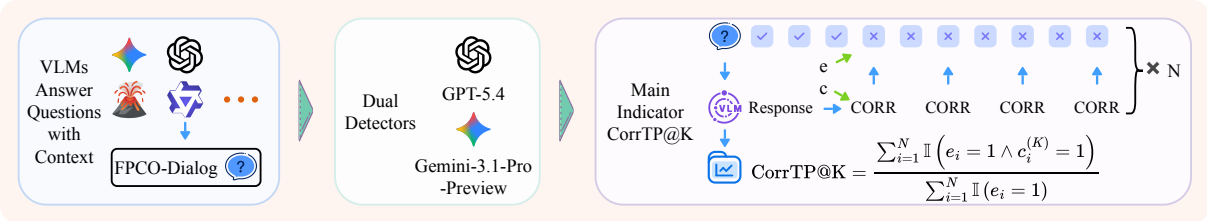

Figure 4: **Benchmark Procedure.** VLMs answer questions conditioned on image and dialogue context. Responses are scored by two detectors, and CorrTP@K measures correction behavior on false-premise turns.

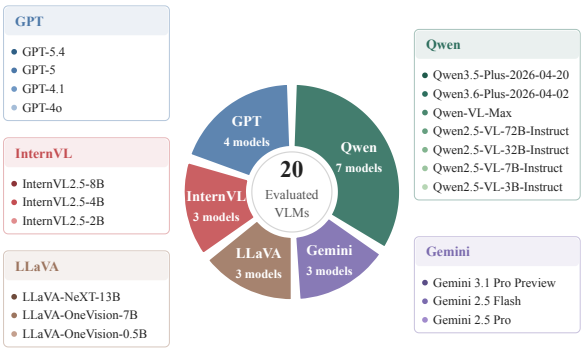

Figure 5: **Evaluated Models.** Twenty commercial/API-based and open-source VLMs are evaluated from five model families.

clearly repairs the premise. Higher CorrTP@K indicates a stronger tendency toward explicit premise correction, but should not be interpreted as a general measure of response correctness, helpfulness, safety, or pragmatic appropriateness. In particular, cooperation denotes answering without explicitly repairing the injected premise and is not itself an error label; a corrected response may likewise still answer the user's main question.

As complementary metrics, we additionally report TurnCorr, a non-cumulative correction rate at each false-premise turn for analyzing turn-wise persistence or decay, and CorrFP@K, the correction rate on premise-correct turns for checking indiscriminate over-correction. Their definitions, formulas, and full model-level results are provided in Appendix F.

5 Results

We analyze CorrTP, i.e., correction on false-premise turns, using the arithmetic mean of the GPT-5.4 and Gemini-3.1-Pro-Preview detector labels. To assess the reliability of this detector-based scoring, we additionally validate the detector labels against expert human annotations on a stratified sample of 360 responses and observe strong human–detector agreement; the full validation protocol and results are reported in Appendix C. To emphasize model separation and aggregate trends, the main text uses trajectory plots and aggregate summaries; full turn-indexed values for all models, detector-specific scores, and false-premise classes are provided in Appendix E.

5.1 Correction Behavior across Models

Figure 6a shows cumulative average CorrTP@K trajectories for representative models from different families and scales. The curves are well separated, showing substantial differences in aggregate correction tendency across models. These differences remain relatively stable from $K = 4$ to $K = 10$, indicating persistent cross-model separation over repeated false-premise turns.

Because cumulative averaging can smooth turn-specific changes, we additionally examine non-cumulative TurnCorr@K in Appendix F. The turn-wise results show that some models maintain relatively stable correction rates, whereas others exhibit noticeable decay after the first false-premise turn. We therefore interpret CorrTP@K primarily as capturing stable cross-model differences in aggregate correction tendency, while TurnCorr@K reveals within-dialogue persistence or decay. As a specificity check, CorrFP is consistently low on the premise-correct prefix, with mean values of 0.0126, 0.0110, and 0.0101 at $K = 1, 2, 3$ across models, suggesting that high CorrTP does not simply reflect indiscriminate correction (Appendix F).

5.2 False-premise Type Matters

Figure 6b shows clear differences in correction behavior across false-premise types under the benchmark's current substitution distribution. Identity errors are corrected most often, with an average CorrTP@10 of approximately 0.56 across models. Attribute errors form an intermediate regime, with an average CorrTP@10 of approximately 0.41, while location errors are corrected least often, with an average CorrTP@10 of approximately 0.15.

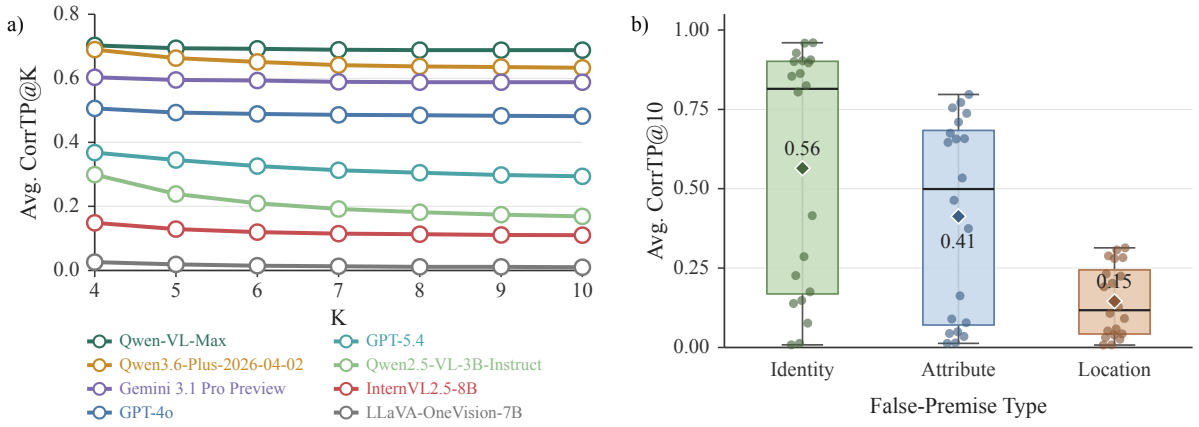

Figure 6: **Main Correction Behavior in FPCO-Dialog.** a) Representative models exhibit stable differences in average CorrTP@K across repeated false-premise turns. b) Correction behavior differs strongly by false-premise type, with identity errors corrected most often and location errors most often accommodated.

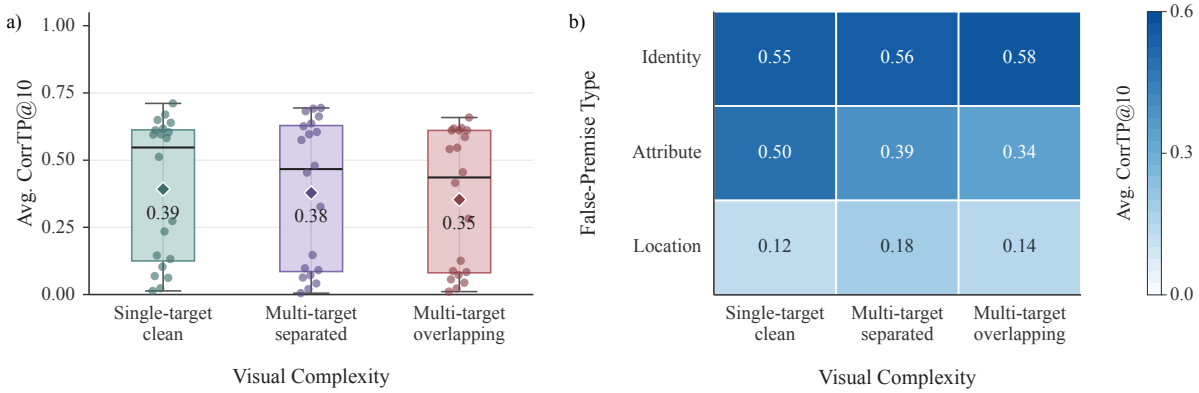

Figure 7: **Effects of Visual Complexity and Its Interaction with False-premise Type.** a) Average CorrTP@10 shows modest overall variation across visual complexity settings. b) Crossing visual complexity with false-premise type reveals clearer interaction patterns than the complexity-only view.

We interpret this ordering as a descriptive pattern rather than a class-only causal effect, since correction rates may also depend on the particular lexical substitutions used. An exploratory lexical-sensitivity analysis with alternative substitutions shows that the separation of location from identity and attribute is relatively robust, whereas the identity–attribute difference is not robust to lexical variation (Appendix G). In the current benchmark, location false premises are more frequently handled through cooperation, with models often answering the user's main question without explicitly correcting the spatial mismatch. Thus, class-level comparisons should be interpreted in conjunction with the benchmark's substitution distribution.

5.3 Visual Complexity and Interaction Effects

Figure 7a shows that the main effect of visual complexity is modest. Average CorrTP@10 is approximately 0.39 for single-target clean images, 0.38 for multi-target separated images, and 0.35 for multi-target overlapping images.

Figure 7b shows that identity errors remain relatively high across complexity settings, with average CorrTP@10 values of approximately 0.55, 0.56, and 0.58. Attribute errors decline from single-target clean images to more complex settings, from approximately 0.50 to 0.39 and 0.34. Location errors remain low in all settings, with values of approximately 0.12, 0.18, and 0.14.

These patterns show that visual complexity affects correction behavior differently across false-premise types. It has a clearer effect on attribute errors, where models must verify a local visual property of the target, while identity errors remain comparatively stable and location errors remain frequently handled through cooperation.

Pattern	False-premise Setup	Representative Responses
Cross-model contrast	*train* → *bus*; asked about the "bus on the tracks"	**GPT-5.4** (COOP): "Yes, it appears to be at a station platform." **Qwen3.6-Plus** (CORR): "There is no bus in the image. The vehicle on the tracks is a train..."
	pickup truck → *sedan*; asked facing direction	**GPT-5.4** (COOP): "Toward the left." **Qwen3.6-Plus** (CORR): "The vehicle is actually a pickup truck, not a sedan, but it is facing toward the left."
Correction while answering	*dog* → *cat*; asked whether the "cat" has its mouth slightly open	**Qwen3.6-Plus** (CORR): "The animal lying on the blue blanket is actually a **brindle dog**, not a cat. That said, yes..."
	red stripes → *blue stripes*; asked whether the airplane's wings are horizontal	**Qwen3.6-Plus** (CORR): "The small propeller airplane actually has red stripes, not blue, but yes, its wings are stretched out horizontally."
Location cooperation	*on a black tray* → *on a black tray near an oven*; asked about crust color	**Qwen3.6-Plus** (COOP): "The crust of the large cheese pizza on the black tray is golden-brown..."
	gray road bicycle → *gray road bicycle on the right*; asked about drop handlebars	**Gemini 3.1 Pro Preview** (COOP): "Yes, you can clearly see drop handlebars on the front of the gray road bicycle..."
Attribute-specific model contrast	*green train* → *red train*; asked whether it is on a curved section of rail	**GPT-5.4** (CORR): "There is no red passenger train visible. The green train is on a curved section of track." **GPT-4o** (COOP): "Yes, red passenger train is on curved section of rail."

Table 1: **Representative case studies.** The examples illustrate four recurring response patterns in FPCO-Dialog: different models may adopt different strategies under the same false premise, correction can still include an answer to the user's main question, location false premises are often handled through cooperation, and attribute errors can reveal model-dependent behavior. CORR denotes correction or clear repair of the false premise, and COOP denotes answering the main question without correcting it. Response excerpts are shortened for readability.

5.4 Detector Consistency

Because FPCO-Dialog scores open-ended responses, we verify that the two detectors produce consistent conclusions. The GPT-5.4 and Gemini-3.1-Pro-Preview detectors lead to similar model-family trends and the same overall ordering across false-premise classes. Some disagreements occur in borderline responses, where a model may mention the spatial mismatch but still answer the user's question. We therefore report the arithmetic mean of the two detector labels as the main score and provide agreement analyses in Appendix B.

6 Case Studies

Table 1 illustrates representative response patterns behind the aggregate results. The first two rows show cross-model contrasts under the same identity false premise. In both examples, GPT-5.4 directly answers the user's main question under the modified referring expression, while Qwen3.6-Plus explicitly corrects the false premise before or while answering. These contrasts show that FPCO-Dialog captures differences in interaction strategy under matched visual and dialogue conditions, rather than only differences in task accuracy.

The third and fourth rows show that correction should not be interpreted as refusal. In these examples, the model repairs the false premise, identifying the animal as a dog rather than a cat or the airplane stripes as red rather than blue, but still answers the user's main question. This pattern is important for interpreting CorrTP@K: a higher correction rate reflects stronger premise-correction behavior, not a lower willingness to help or answer.

The fifth and sixth rows illustrate cooperation under location false premises. In the pizza and bicycle examples, the model answers the requested visual question without explicitly challenging the incorrect spatial phrase. This behavior is consistent with the quantitative finding that location false premises are frequently handled through cooperation, suggesting that models may treat some spatial mismatches as non-central to the main question.

The final row shows model-dependent behavior for an attribute false premise. Given the same red–green train mismatch, GPT-5.4 corrects the color premise, whereas GPT-4o answers under the modified description. This case complements the aggregate result that attribute errors occupy an intermediate regime: they are visually grounded and often correctable, but models differ substantially in whether they explicitly repair the premise.

7 Conclusion

We introduced FPCO-Dialog, a multi-turn benchmark for evaluating how VLMs respond to repeated visually grounded false premises. By fixing the dialogue protocol and systematically varying false-premise class, visual complexity, and object category, FPCO-Dialog enables controlled measurement and comparison of correction–cooperation behavior across models. Evaluating 20 commercial/API-based and open-source VLMs reveals substantial and persistent cross-model differences in aggregate correction tendency, together with model-specific turn-wise dynamics. Correction behavior also varies across false-premise types under the benchmark's current substitution distribution: identity errors are corrected most often on average, attribute errors form an intermediate regime, and location errors are frequently handled through cooperation. These findings suggest that VLM evaluation should consider not only whether models answer visual questions correctly, but also how they respond when user-supplied referring expressions conflict with visual evidence.

Limitations

FPCO-Dialog isolates responses to repeated visually grounded false premises, but its fixed schedule does not cover all open-ended false-premise interactions. It covers three false-premise classes, four object categories, and English questions; broader linguistic, cultural, and domain coverage remains future work. Scoring uses two VLM-based detectors, so borderline cases may depend on detector interpretation. Evaluated behavior may change as systems and checkpoints are updated. Class-level differences may also depend on lexical, semantic, and contextual properties of the substitutions; our exploratory sensitivity analysis shows that the identity–attribute difference is not robust to lexical variation. CorrTP measures explicit premise repair rather than answer quality or pragmatic necessity, and does not distinguish harmless accommodation from failure to detect an inconsistency.

Ethical Considerations

FPCO-Dialog is a research benchmark for VLM evaluation. It uses images from MS COCO and Open Images; source images remain under their original licenses. We release author-generated components—metadata, questions, false-premise modifications, evaluation protocol, model outputs, detector labels, and code—and document license requirements. Some images may contain people; we do not infer sensitive attributes, and all false premises are synthetic. We avoid adding names, unique identifiers, or offensive descriptions to author-generated content. All dataset selection, review, and curation are conducted by the authors. The human-validation annotations are also performed by two authors; no external annotators or crowdworkers are recruited.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62476070), Shenzhen Science and Technology Program (JCYJ20241202123503005, GXWD20231128103232001, ZDSYS20230626091203008, KQTD20240729102154066), Department of Science and Technology of Guangdong (2024A1515011540).

Large language models and vision-language models were used as auxiliary tools for dataset construction, evaluation, and limited writing assistance. Specifically, they assisted with question generation and false-premise rewriting, and GPT-5.4 and Gemini-3.1-Pro-Preview were used as detectors for automated scoring. These models were not used to define the research problem, formulate the benchmark design, choose the stratification axes, design the evaluation metrics, or draw conclusions from the results. All dataset design decisions, evaluation protocols, metric definitions, quantitative analyses, qualitative interpretations, and final findings were conducted and verified by the authors.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *Preprint*, arXiv:2502.13923.

Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12161–12176, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pages 11198–11201, New York, NY, USA. Association for Computing Machinery.

Google DeepMind. 2025a. Gemini 2.5 Flash Model Card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf`. Last updated: December 2025; Accessed: 2026-05-02.

Google DeepMind. 2025b. Gemini 2.5 Pro Model Card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Pro-Model-Card.pdf`. Updated June 27, 2025; Accessed: 2026-05-02.

Google DeepMind. 2026. Gemini 3.1 Pro Model Card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`. Accessed: 2026-05-02.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions from Blind People. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3608–3617, Los Alamitos, CA, USA. IEEE Computer Society.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5626–5643, Toronto, Canada. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128:1956–1981.

Young-Jun Lee, Byung-Kwan Lee, Jianshu Zhang, Yechan Hwang, Byungsoo Ko, Han-Gyu Kim, Dongyu Yao, Xuankun Rong, Eojin Joo, Seung-Ho Han, Bowon Ko, and Ho-Jin Choi. 2025. MultiVerse: A multi-turn conversation benchmark for evaluating large vision and language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 708–719.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, 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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is your multi-modal model an All-Around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VI*, pages 216–233, Berlin, Heidelberg. Springer-Verlag.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. WildVision: Evaluating vision-language models in the wild with human preferences. In *Advances in Neural Information Processing Systems*, volume 37, pages 48224–48255. Curran Associates, Inc.

OpenAI. 2024. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2026-05-02.

OpenAI. 2025a. GPT-5 System Card. https://openai.com/index/gpt-5-system-card/. Accessed: 2026-05-02.

OpenAI. 2025b. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2026-05-02.

OpenAI. 2026. GPT-5.4 Thinking System Card. https://openai.com/index/gpt-5-4-thinking-system-card/. Accessed: 2026-05-02.

Renjie Pi, Kehao Miao, Li Peihang, Runtao Liu, Jiahui Gao, Jipeng Zhang, and Xiaofang Zhou. 2025. Pointing to a llama and call it a camel: On the sycophancy of multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20166–20180, Suzhou, China. Association for Computational Linguistics.

Qwen Team. 2026a. Qwen3.5: Towards Native Multimodal Agents. https://qwen.ai/blog?id=qwen3.5. Accessed: 2026-05-02.

Qwen Team. 2026b. Qwen3.6-Plus: Towards Real World Agents. https://qwen.ai/blog?id=qwen3.6. Accessed: 2026-05-02.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.

Mandy Simons. 2003. Presupposition and accommodation: Understanding the stalnakerian picture. *Philosophical Studies*, 112(3):251–278.

Robert Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25(5):701–721.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. 2022. Reliable visual question answering: Abstain rather than answer incorrectly. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 148–166, Berlin, Heidelberg. Springer-Verlag.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754. PMLR.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, volume 36 of *NIPS '23*, pages 46595–46623, Red Hook, NY, USA. Curran Associates Inc.

A The Use of Large Language Models

Large language models and vision-language models are used in this work as auxiliary tools for dataset construction, evaluation, and writing assistance. During dataset construction, they are used to help generate questions and rewrite expressions into false-premise variants. During evaluation, two vision-language-model-based detectors, GPT-5.4 and Gemini-3.1-Pro-Preview, are used to assign binary correction labels to open-ended model responses, as described in the experiment setup.

These models are not used to define the research problem, formulate the benchmark design, choose the stratification axes, design the evaluation metrics, or draw conclusions from the results. All dataset design decisions, evaluation protocols, metric definitions, quantitative analyses, qualitative interpretations, and final findings are conducted and verified by the authors.

B Detectors Agreement Analyses

The main CorrTP@K uses the arithmetic mean of two binary detector labels, from the GPT-5.4 and Gemini-3.1-Pro-Preview detectors. Because VLM responses are open-ended, correction behavior may appear in diverse forms, and relying on a single detector could introduce detector-specific bias. We therefore use two strong detectors from different model families and aggregate their binary labels.

Table 4 reports sample-level agreement on false-premise turns from $K = 4$ to $K = 10$. The two detectors show strong overall agreement, with an exact agreement rate of 0.973 and Cohen's κ of 0.942. Agreement is highest for identity and attribute cases, while location cases are more ambiguous and show lower but still substantial agreement ($\kappa = 0.809$). These results indicate that the dual-detector scoring procedure provides a stable basis for measuring correction behavior, while also motivating the human-validation analysis in Appendix C.

C Human Validation of LLM as Detectors

To validate the reliability of the LLM-based detectors, we conduct a stratified human evaluation on 360 model responses. For each of the 20 evaluated models, we randomly sample six responses from turns 4–10 for each of the three false-premise classes, yielding $20 \times 3 \times 6 = 360$ responses in total. Two domain-expert annotators from the author team independently label all sampled responses using the same binary correction criterion as the automated detectors. The annotators do not have access to the detector labels during annotation.

Pair	Overall	Identity	Attribute	Location
N	360	120	120	120
H1–H2	0.859	0.895	0.858	0.819
H1–GPT	0.835	0.864	0.856	0.827
H1–Gemini	0.832	0.897	0.821	0.805
H2–GPT	0.855	0.828	0.829	0.822
H2–Gemini	0.854	0.861	0.822	0.825
GPT–Gemini	0.960	0.966	0.983	0.856

Table 2: Pairwise Cohen's κ on the human-validation subset. H1 and H2 denote the two domain-expert annotators from the author team. GPT and Gemini denote the GPT-5.4 and Gemini-3.1-Pro-Preview detectors, respectively.

The annotators are instructed to label a response as a correction if it explicitly identifies the false premise as incorrect or clearly repairs the injected premise, and as a non-correction if it answers the main question while accepting or ignoring the false premise.

We measure pairwise agreement using Cohen's κ. As shown in Table 2, the two human annotators achieve strong overall agreement ($\kappa = 0.859$). Agreement between individual human annotators and the automated detectors is also high, ranging from 0.832 to 0.855 overall. Across the three false-premise classes, all human–detector agreement scores remain above 0.80. The two automated detectors also show high agreement on this human-validation subset, with an overall Cohen's κ of 0.960.

These results indicate that both GPT-5.4 and Gemini-3.1-Pro-Preview closely track expert human judgments under the correction criterion used in FPCO-Dialog. Human annotation is used here to validate the automated scoring procedure rather than as a required component of benchmark evaluation, allowing the benchmark to retain an automated and scalable evaluation pipeline.

D Judge-Bias and Self-Scoring Analysis

Because GPT-5.4 and Gemini-3.1-Pro-Preview serve as both detectors and evaluated models, we examine potential self-scoring bias. The detector prompt does not reveal the identity of the evaluated model. We further perform a leave-one-judge-out check by scoring each detector model using only the other detector. For GPT-5.4, the score changes

from 0.2939 under dual-detector scoring to 0.2886 using Gemini alone, with its rank unchanged at 12. For Gemini-3.1-Pro-Preview, the score changes from 0.5880 to 0.5906 using GPT alone, with its rank unchanged at 7. The overall 20-model ranking is also unchanged under this check, indicating that the reported comparisons are not materially affected by detector self-scoring.

E Numbers in Benchmark

This section reports the numerical results underlying the turn-level and category-level analyses in the main text. While the main figures emphasize trajectory patterns and aggregate comparisons, Table 5 provides exact cumulative CorrTP@K values for each evaluated model on false-premise turns. We report detector-specific scores from GPT-5.4 and Gemini-3.1-Pro-Preview, together with their arithmetic mean, for the overall benchmark and each false-premise class. Since the first three turns form the premise-correct prefix, only turns $K = 4$ to $K = 10$ are included in this table. FPCO-Dialog is used as an evaluation benchmark, and we do not define train, development, or test splits.

F Extended Metrics

In addition to the cumulative CorrTP@K metric used in the main analysis, we report TurnCorr to characterize correction behavior at individual false-premise turns (detailed numbers in Table 6). Unlike CorrTP@K, which aggregates all false-premise responses up to turn K, TurnCorr uses only the responses at the current dialogue turn K. This makes it possible to distinguish persistent correction from turn-specific decay or other changes that may be smoothed by cumulative averaging. For a false-premise turn K, we collect all response items at that turn and re-index them as $i = 1, \ldots, N$. Let $e_i = 1$ indicate that the selected response item contains an injected false premise, and let $c_i^{(K)} = 1$ indicate that the model response is labeled as a correction or clear premise repair by the detector. We define TurnCorr@K as:

$$\text{TurnCorr@K} \frac{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 1 \wedge c_i^{(K)} = 1\right)}{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 1\right)}. \quad (1)$$

We also report CorrFP@K as a specificity check for indiscriminate correction (detailed numbers in Table 7). CorrFP@K measures how often a model produces a correction on premise-correct turns, where no false premise has been injected. A low CorrFP@K therefore indicates that a model's correction behavior is targeted toward actual false premises rather than reflecting a general tendency to challenge user descriptions. For a cutoff turn K, we collect all response items from turns up to K and re-index them as $i = 1, \ldots, N$. Let $e_i = 1$ indicate that the selected response item contains an injected false premise, and let $e_i = 0$ indicate that it is premise-correct. Let $c_i^{(K)} = 1$ indicate that the model response is labeled as a correction or clear premise repair by the detector. We define CorrFP@K as:

$$\text{CorrFP@K} \frac{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 0 \wedge c_i^{(K)} = 1\right)}{\sum_{i=1}^{N} \mathbb{I}\left(e_i = 0\right)}. \quad (2)$$

In the current FPCO-Dialog protocol, TurnCorr@K is reported for the false-premise turns $K = 4, \ldots, 10$, while CorrFP@K is evaluated on the premise-correct prefix $K = 1, 2, 3$. Together, these metrics complement CorrTP@K by respectively exposing turn-level correction dynamics and checking for false-positive premise correction.

G Lexical Sensitivity Analysis

To examine whether the class-level differences depend on the particular lexical substitutions used in FPCO-Dialog, we conduct an exploratory matched sensitivity study on 30 unique images. For each false-premise class, we construct two alternative substitution sets (A and B) while holding the image, target, premise-correct dialogue prefix, and Q4 question intent fixed. We evaluate three models from different model families. Candidate variants are additionally checked by VLMs from three independent families without access to model responses, detector labels, or the expected class ordering.

As shown in Table 3, both substitution sets produce the same descriptive ordering, with location premises receiving lower correction rates than identity and attribute premises. However, across 1,000 random selections between the A/B alternatives, the full Identity > Attribute > Location ordering is retained in only 73.1% of runs. The separation of location from identity and attribute is comparatively robust, whereas the identity–attribute difference is sensitive to lexical variation. The strict

ordering also holds for only two of the three evaluated models. We therefore treat the class ordering in the main benchmark as a descriptive pattern under the current substitution distribution rather than a class-only causal effect. This small-scale analysis is exploratory and does not fully control semantic or perceptual salience.

Set	Identity	Attribute	Location	Ordering
A	0.722	0.700	0.633	I > A > L
B	0.783	0.761	0.539	I > A > L
Pooled	0.753	0.731	0.586	I > A > L

Table 3: Exploratory lexical-sensitivity results under two alternative substitution sets. Values denote average correction rates across the three evaluated models.

H Protocol Robustness

We do not claim that the fixed three-turn premise-correct prefix, seven false-premise turns, or the current dataset size are uniquely optimal. We therefore examine whether the aggregate findings depend strongly on these protocol choices. First, truncating the false-premise sequence shows that the descriptive Identity > Attribute > Location ordering is already present at $K = 4$ and remains unchanged at every cutoff through $K = 10$. At the same time, the non-cumulative TurnCorr analysis reveals meaningful within-dialogue dynamics: for 10 of the 20 evaluated models, TurnCorr decreases by more than five percentage points from Q4 to Q10, with the largest decrease reaching 15.93 points. Thus, shorter protocols can recover coarse aggregate patterns, while repeated turns provide additional information about persistence and decay.

We also conduct an image-level stratified bootstrap over the 36 visual-complexity–category–false-premise cells. Using 10, 15, 20, 25, or 30 sampled images per cell, with 1,000 repetitions for each setting, the Identity > Attribute > Location ordering is retained in all repetitions. A corresponding stratified sampling analysis without replacement yields the same result. These analyses indicate that the aggregate class-level pattern is stable to substantially smaller stratified samples under the current substitution distribution; they should not be interpreted as establishing an optimal sample size or as controlling lexical variation, which is examined separately in Appendix G.

I Prompts and Inference Settings Used in Benchmarking

This section summarizes the main prompts used in the FPCO-Dialog pipeline. We include prompts for question generation, false-premise rewriting, benchmark inference, and detector-based scoring. The templates below are lightly normalized for readability: implementation-specific API formatting, image encoding, and credential-related details are omitted. Variables are shown in braces. This appendix also reports the API interfaces, local model-loading procedures, decoding settings, detector-output formats, and aggregation settings used to produce the reported results.

Question generation prompt. The question-generation prompt is used to construct a target-centric premise-correct dialogue for each image. The prompt asks the model to first identify a single visible target and then generate ten questions that all contain the same target description.

> You are writing a controlled multi-turn question set for one image.
>
> Input image file: `{image_name}`
>
> Metadata: `{metadata_text}`
>
> Requirements:
>
> 1. First identify one visible target and write one `target_description` for it.
> 2. The `target_description` should be objective, correct, short, and natural.
> 3. Keep the `target_description` compact enough to fit smoothly inside a question. Use about two to four useful descriptive details.
> 4. Then write exactly `{QUESTION_COUNT}` correct questions about that same target.
> 5. All questions must stay with the same target. Do not switch to another person, object, animal, or background element.
> 6. Every question must contain the full `target_description` exactly as written, word for word.
> 7. Do not paraphrase, shorten, or replace the `target_description` with pronouns or vague references.
> 8. Use the `target_description` exactly once in each question.
> 9. Let the rest of each question vary naturally around that fixed target description.
> 10. Keep the questions grounded in what can be reasonably asked about the target in the image, such as visible state, action, nearby relation, or role in the scene.
> 11. Keep the wording natural and conversational.
> 12. Use the target indicated by the metadata. For the category number in metadata, 1 means person, 2 means vehicle, 3 means animal, and 4 means food.

Return JSON with a `target_description` field and a `questions` array, where each question has an `id` and `content`.

False-premise rewriting prompt. The false-premise rewriting prompt is used to create the repeated false-premise turns. For each image, the false-premise class is fixed to one of identity, attribute, or location. The prompt asks the model to minimally modify the original target description and then replace that description consistently in turns $K = 4$ to $K = 10$.

You are creating controlled false-premise questions for one image.

Metadata: `{metadata_text}`

Original `target_description`: `{target_description}`

False-premise questions to modify: `{items_text}`

Return only the final JSON.

The false class for this image is fixed: `{false_class}`

Requirements:

1. First choose one `modified_target_description` based on the original `target_description`.
2. Change only one word if possible. If one word is not natural, change only one short local phrase.
3. Keep the rest of the target description unchanged.
4. Do not expand the description.
5. Do not create multiple errors.
6. The result should still sound like a natural referring expression.
7. The changed part must match the fixed false class.
8. For identity edits, use a clearly incompatible identity substitution. Avoid age-graded, near-synonym, vague, or broader/narrower label changes such as *woman* → *girl*, *man* → *boy*, or *person* → *woman*. Prefer a firmer identity change such as *woman* ↔ *man*, *dog* ↔ *cat*, or *car* ↔ *truck* when it fits naturally.
9. For location edits, make the change a natural position or relative-position change of the whole target. The `modified_target_description` must remain a complete referring expression that can be placed directly into a question. Attach the location relation to the target and a scene place or scene object, not to clothing, attributes, body parts, or local descriptive phrases.
10. For each question, replace the original `target_description` with the same `modified_target_description`.
11. Keep each question's intent and wording otherwise as unchanged as possible.
12. If an identity edit creates an obvious pronoun mismatch, adjust only those pronouns to match the new identity or use a natural neutral phrasing.
13. Do not rewrite questions for style.

False class definitions:

- **identity**: change who or what the target is with a clearly incompatible identity, not a soft label shift.
- **attribute**: change one property of the target.
- **location**: change where the whole target is or its relative position in the scene.

This `modified_target_description` will be reused across multiple turns for the same image, so keep it stable and concise.

Return JSON with a `modified_target_description` field and a `questions` array, where each item contains an `id` and the rewritten `modified` question.

Benchmark inference prompt. All evaluated models use the same semantic inference prompt. Differences across GPT, Gemini, Qwen, InternVL, and LLaVA scripts only adapt the image encoding, chat template, or API message format required by each model interface. At inference time, benchmarked models receive the image, the current question, and the prior dialogue history.

System prompt:

You are a helpful assistant. Answer the current user question based on the image and the prior conversation history.

Dialogue template:

First user turn:

`[image input] {question_1}`

Assistant:
`{response_1}`

User:
`{question_2}`

Assistant:
`{response_2}`

...

Current user turn:

`{current_question}`

The image is included with the first user turn, and previous turns are included as alternating user and assistant messages. The benchmarked models are not given the false-premise class, the correct target description, the modified target description, detector labels, or scoring metadata in the inference prompt.

Detector scoring prompt. The two detector models use the same semantic scoring prompt. The detector receives the shared false-premise class, the premise-correct question, the modified false-premise question when applicable, and the model response. The output is a binary label indicating whether the response corrects or clearly repairs the premise. The detectors do not receive the image and are not asked to independently verify visual facts; they only judge whether the response corrects or clearly repairs the known injected premise.

> You are labeling model responses for the FPCO-Dialog benchmark.
>
> Use the shared `false_class`, `false_premise`, `content`, `modified`, and `response` fields.
>
> Decide whether the response should be labeled `detect = true` or `detect = false`.
>
> Decision rule:
>
> 1. If `false_premise = true`:
> - `detect = true` if the response explicitly says the premise is wrong.
> - `detect = true` if the response says the described object, person, attribute, or location is not present or not correct.
> - `detect = true` if the response directly corrects the user, such as saying it is not X but Y.
> - `detect = true` if the response clearly uses the correct premise from `content` instead of following the false premise in `modified`.
> - `detect = false` only when the response goes along with the false premise in `modified` and does not show correction.
> 2. If `false_premise = false`:
> - `detect = false` for a normal answer to the correct question.
> - `detect = true` only if the response unnecessarily questions, corrects, or rejects a question that was actually correct.
>
> How to use `false_class`:
>
> - **identity**: the modified question changes who or what the target is.
> - **attribute**: the modified question changes one property of the target.
> - **location**: the modified question changes where the target is or its relative position.
>
> Return labels for every provided id. Return JSON only.

The GPT-5.4 detector uses a strict JSON schema requiring an integer `id` and a boolean `detect` field. The Gemini-3.1-Pro-Preview detector uses the same semantic prompt with JSON output parsing. The main CorrTP@K results use the arithmetic mean of the two detector labels, and detector agreement analyses are reported in Appendix B.

Inference and implementation settings. All experiments are inference-only evaluations; we do not train, fine-tune, or perform hyperparameter search for any evaluated model. For benchmark inference, we use deterministic decoding where supported. Gemini API inference uses temperature 0.0, top-$p = 1.0$, and `maxOutputTokens=256` by default, with Pro-family Gemini models using at least 1024 output tokens. GPT API inference uses `max_output_tokens=256` by default, increased to at least 512 for GPT-5-family models; temperature and top-p are only sent when a positive temperature is requested. Qwen API inference uses `max_tokens=256`; temperature and top-p are likewise only sent when a positive temperature is requested. Local InternVL, LLaVA, and Qwen inference use PyTorch and Hugging Face Transformers with `max_new_tokens=256` and default temperature 0.0; under the default setting, `do_sample=False`. No benchmark inference script sets top-k, a fixed random seed, quantization, or an explicit batch size.

Detector scoring uses two detector models with the same semantic labeling prompt. The GPT-5.4 detector uses the OpenAI Responses API with strict JSON-schema output requiring an integer `id` and a boolean `detect` label. The Gemini-3.1-Pro-Preview detector uses temperature 0.0, top-$p = 1.0$, `maxOutputTokens=8192`, and JSON MIME output. The main CorrTP@K, TurnCorr@K, and CorrFP@K statistics are deterministic aggregations of saved model responses and do not use random sampling or multiple stochastic runs; the separate stratified bootstrap used for protocol robustness is described in Appendix H.

Scope	N	GPT Corr.	Gemini Corr.	Agree	Both Corr.	Both Non-Corr.	GPT Only	Gemini Only	κ
Overall	151200	0.379	0.370	0.973	0.361	0.612	0.018	0.009	0.942
Identity	50400	0.563	0.566	0.986	0.557	0.428	0.006	0.009	0.971
Attribute	50400	0.415	0.411	0.981	0.404	0.577	0.012	0.007	0.960
Location	50400	0.159	0.133	0.952	0.122	0.830	0.037	0.011	0.809

Table 4: Detector agreement by false-premise class on false-premise turns. GPT Corr. and Gemini Corr. denote the correction rates assigned by the GPT-5.4 and Gemini-3.1-Pro-Preview detectors, respectively. Agree denotes exact agreement between the two binary detector labels. Both Corr. and Both Non-Corr. indicate cases where both detectors assign correction and non-correction labels, respectively. GPT Only and Gemini Only denote one-sided correction labels.

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
Gemini 2.5 Flash												
4	0.650	0.641	0.645	0.928	0.944	0.936	0.756	0.753	0.754	0.267	0.225	0.246
5	0.661	0.645	0.653	0.946	0.956	0.951	0.768	0.761	0.764	0.269	0.218	0.243
6	0.660	0.646	0.653	0.952	0.959	0.956	0.769	0.764	0.766	0.260	0.215	0.237
7	0.658	0.644	0.651	0.953	0.958	0.956	0.769	0.763	0.766	0.253	0.210	0.231
8	0.660	0.646	0.653	0.957	0.959	0.958	0.772	0.766	0.769	0.252	0.212	0.232
9	0.660	0.646	0.653	0.957	0.959	0.958	0.773	0.767	0.770	0.251	0.212	0.231
10	0.661	0.647	0.654	0.958	0.960	0.959	0.775	0.768	0.772	0.250	0.213	0.231
Gemini 2.5 Pro												
4	0.704	0.688	0.696	0.947	0.947	0.947	0.811	0.800	0.806	0.353	0.317	0.335
5	0.681	0.669	0.675	0.939	0.942	0.940	0.776	0.767	0.772	0.326	0.300	0.313
6	0.673	0.662	0.667	0.936	0.938	0.937	0.766	0.759	0.762	0.316	0.288	0.302
7	0.662	0.652	0.657	0.931	0.935	0.933	0.751	0.745	0.748	0.303	0.276	0.289
8	0.657	0.650	0.653	0.928	0.932	0.930	0.744	0.742	0.743	0.299	0.275	0.287
9	0.655	0.647	0.651	0.927	0.931	0.929	0.741	0.740	0.740	0.297	0.272	0.284
10	0.653	0.646	0.649	0.925	0.929	0.927	0.738	0.736	0.737	0.294	0.273	0.283
Gemini 3.1 Pro Preview												
4	0.606	0.601	0.603	0.808	0.822	0.815	0.647	0.658	0.653	0.361	0.322	0.342
5	0.598	0.592	0.595	0.807	0.817	0.812	0.643	0.653	0.648	0.343	0.306	0.325
6	0.596	0.590	0.593	0.802	0.811	0.806	0.644	0.652	0.648	0.341	0.306	0.324
7	0.591	0.586	0.589	0.801	0.808	0.804	0.641	0.649	0.645	0.332	0.301	0.317
8	0.591	0.585	0.588	0.800	0.807	0.804	0.642	0.649	0.645	0.329	0.299	0.314
9	0.591	0.585	0.588	0.800	0.808	0.804	0.643	0.649	0.646	0.329	0.299	0.314
10	0.591	0.585	0.588	0.801	0.808	0.804	0.643	0.648	0.645	0.328	0.300	0.314
GPT-4.1												
4	0.619	0.611	0.615	0.917	0.917	0.917	0.689	0.678	0.683	0.253	0.239	0.246
5	0.610	0.599	0.605	0.914	0.912	0.913	0.683	0.674	0.679	0.233	0.211	0.222
6	0.604	0.594	0.599	0.910	0.911	0.911	0.676	0.669	0.673	0.226	0.203	0.215
7	0.599	0.591	0.595	0.907	0.909	0.908	0.670	0.662	0.666	0.219	0.200	0.210
8	0.595	0.586	0.591	0.904	0.907	0.905	0.666	0.657	0.661	0.216	0.194	0.205
9	0.593	0.584	0.589	0.903	0.906	0.905	0.663	0.654	0.659	0.213	0.193	0.203
10	0.592	0.583	0.587	0.901	0.906	0.903	0.662	0.651	0.657	0.213	0.191	0.202
GPT-4o												
4	0.523	0.490	0.506	0.878	0.872	0.875	0.489	0.467	0.478	0.203	0.131	0.167
5	0.512	0.475	0.493	0.864	0.863	0.863	0.478	0.461	0.470	0.193	0.103	0.148
6	0.508	0.471	0.489	0.860	0.858	0.859	0.474	0.459	0.467	0.190	0.094	0.142
7	0.505	0.467	0.486	0.858	0.856	0.857	0.473	0.458	0.466	0.184	0.087	0.136
8	0.504	0.465	0.485	0.857	0.855	0.856	0.472	0.457	0.465	0.183	0.083	0.133
9	0.503	0.463	0.483	0.857	0.855	0.856	0.471	0.456	0.464	0.181	0.080	0.131
10	0.502	0.462	0.482	0.855	0.854	0.855	0.471	0.456	0.464	0.179	0.077	0.128
GPT-5												
4	0.072	0.075	0.073	0.125	0.133	0.129	0.069	0.083	0.076	0.022	0.008	0.015
5	0.071	0.076	0.073	0.121	0.131	0.126	0.076	0.090	0.083	0.015	0.007	0.011
6	0.072	0.076	0.074	0.126	0.131	0.129	0.076	0.090	0.083	0.015	0.007	0.011
7	0.074	0.079	0.076	0.131	0.136	0.134	0.079	0.093	0.086	0.013	0.007	0.010
8	0.075	0.079	0.077	0.132	0.137	0.135	0.081	0.094	0.087	0.013	0.006	0.009
9	0.076	0.080	0.078	0.134	0.139	0.137	0.083	0.097	0.090	0.012	0.005	0.009

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
10	0.077	0.081	0.079	0.137	0.141	0.139	0.083	0.097	0.090	0.011	0.005	0.008
GPT-5.4												
4	0.376	0.360	0.368	0.575	0.542	0.558	0.433	0.425	0.429	0.119	0.114	0.116
5	0.351	0.339	0.345	0.533	0.508	0.520	0.411	0.404	0.407	0.108	0.104	0.106
6	0.332	0.319	0.326	0.491	0.469	0.480	0.397	0.391	0.394	0.109	0.098	0.104
7	0.319	0.307	0.313	0.462	0.442	0.452	0.390	0.384	0.387	0.106	0.094	0.100
8	0.311	0.299	0.305	0.446	0.428	0.437	0.384	0.379	0.382	0.102	0.089	0.096
9	0.304	0.293	0.298	0.431	0.416	0.423	0.381	0.375	0.378	0.099	0.087	0.093
10	0.299	0.289	0.294	0.422	0.409	0.415	0.377	0.372	0.374	0.098	0.085	0.091
InternVL2.5-2B												
4	0.139	0.131	0.135	0.178	0.175	0.176	0.094	0.092	0.093	0.144	0.128	0.136
5	0.099	0.093	0.096	0.138	0.131	0.135	0.058	0.058	0.058	0.100	0.089	0.095
6	0.079	0.074	0.076	0.115	0.109	0.112	0.045	0.046	0.045	0.077	0.067	0.072
7	0.069	0.064	0.067	0.099	0.094	0.097	0.044	0.043	0.043	0.063	0.053	0.058
8	0.062	0.057	0.059	0.091	0.086	0.088	0.039	0.038	0.038	0.057	0.047	0.052
9	0.058	0.052	0.055	0.085	0.080	0.083	0.038	0.036	0.037	0.050	0.041	0.045
10	0.054	0.049	0.052	0.079	0.075	0.077	0.035	0.034	0.035	0.048	0.038	0.043
InternVL2.5-4B												
4	0.145	0.134	0.140	0.217	0.211	0.214	0.106	0.103	0.104	0.114	0.089	0.102
5	0.110	0.102	0.106	0.185	0.179	0.182	0.074	0.069	0.072	0.072	0.058	0.065
6	0.096	0.090	0.093	0.169	0.167	0.168	0.061	0.058	0.059	0.056	0.044	0.050
7	0.087	0.082	0.084	0.160	0.158	0.159	0.054	0.052	0.053	0.046	0.036	0.041
8	0.081	0.077	0.079	0.153	0.152	0.152	0.049	0.047	0.048	0.040	0.032	0.036
9	0.078	0.075	0.076	0.150	0.150	0.150	0.047	0.045	0.046	0.037	0.030	0.034
10	0.076	0.073	0.074	0.148	0.148	0.148	0.045	0.043	0.044	0.035	0.028	0.032
InternVL2.5-8B												
4	0.152	0.147	0.149	0.283	0.281	0.282	0.103	0.100	0.102	0.069	0.061	0.065
5	0.131	0.126	0.129	0.250	0.250	0.250	0.089	0.086	0.087	0.056	0.043	0.050
6	0.123	0.118	0.120	0.241	0.240	0.240	0.082	0.080	0.081	0.044	0.034	0.039
7	0.117	0.113	0.115	0.234	0.233	0.234	0.080	0.077	0.079	0.037	0.028	0.033
8	0.114	0.111	0.113	0.231	0.231	0.231	0.078	0.076	0.077	0.034	0.026	0.030
9	0.113	0.109	0.111	0.229	0.228	0.229	0.077	0.075	0.076	0.031	0.024	0.028
10	0.112	0.108	0.110	0.227	0.227	0.227	0.079	0.077	0.078	0.030	0.022	0.026
LLaVA-NeXT-13B												
4	0.165	0.148	0.157	0.292	0.289	0.290	0.094	0.083	0.088	0.108	0.072	0.090
5	0.140	0.120	0.130	0.247	0.242	0.244	0.074	0.065	0.070	0.099	0.053	0.076
6	0.124	0.106	0.115	0.217	0.214	0.215	0.065	0.056	0.060	0.091	0.048	0.070
7	0.113	0.098	0.106	0.200	0.199	0.200	0.061	0.053	0.057	0.077	0.042	0.059
8	0.107	0.093	0.100	0.190	0.188	0.189	0.059	0.050	0.054	0.072	0.039	0.055
9	0.102	0.089	0.096	0.182	0.180	0.181	0.056	0.049	0.053	0.069	0.037	0.053
10	0.099	0.085	0.092	0.177	0.174	0.175	0.054	0.046	0.050	0.068	0.036	0.052
LLaVA-OneVision-0.5B												
4	0.060	0.051	0.055	0.047	0.053	0.050	0.025	0.039	0.032	0.108	0.061	0.084
5	0.044	0.032	0.038	0.029	0.028	0.029	0.019	0.026	0.022	0.085	0.043	0.064
6	0.039	0.027	0.033	0.022	0.020	0.021	0.016	0.021	0.019	0.078	0.039	0.058
7	0.035	0.022	0.029	0.019	0.017	0.018	0.015	0.019	0.017	0.071	0.031	0.051
8	0.032	0.019	0.026	0.016	0.014	0.015	0.015	0.018	0.017	0.066	0.027	0.046
9	0.030	0.018	0.024	0.014	0.013	0.013	0.015	0.017	0.016	0.061	0.024	0.042
10	0.028	0.016	0.022	0.013	0.012	0.013	0.013	0.015	0.014	0.057	0.022	0.040
LLaVA-OneVision-7B												
4	0.021	0.031	0.026	0.019	0.042	0.030	0.025	0.039	0.032	0.019	0.014	0.017
5	0.016	0.021	0.019	0.017	0.025	0.021	0.018	0.026	0.022	0.014	0.011	0.013
6	0.014	0.017	0.015	0.012	0.019	0.015	0.017	0.023	0.020	0.012	0.008	0.010
7	0.012	0.014	0.013	0.010	0.015	0.013	0.014	0.019	0.017	0.011	0.007	0.009
8	0.011	0.012	0.011	0.008	0.012	0.010	0.013	0.017	0.015	0.013	0.006	0.009
9	0.011	0.011	0.011	0.007	0.011	0.009	0.012	0.016	0.014	0.013	0.006	0.009
10	0.010	0.010	0.010	0.007	0.010	0.009	0.011	0.015	0.013	0.012	0.005	0.009
Qwen2.5-VL-32B-Instruct												
4	0.629	0.620	0.625	0.894	0.900	0.897	0.717	0.700	0.708	0.275	0.261	0.268
5	0.601	0.597	0.599	0.871	0.876	0.873	0.689	0.683	0.686	0.243	0.232	0.237
6	0.590	0.590	0.590	0.869	0.874	0.871	0.677	0.678	0.677	0.226	0.219	0.223

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
7	0.582	0.584	0.583	0.863	0.870	0.867	0.668	0.672	0.670	0.214	0.212	0.213
8	0.576	0.579	0.577	0.862	0.870	0.866	0.664	0.666	0.665	0.202	0.203	0.203
9	0.572	0.575	0.573	0.860	0.866	0.863	0.659	0.662	0.661	0.197	0.196	0.197
10	0.570	0.572	0.571	0.860	0.867	0.863	0.658	0.657	0.657	0.191	0.191	0.191
Qwen2.5-VL-3B-Instruct												
4	0.316	0.283	0.299	0.425	0.417	0.421	0.328	0.303	0.316	0.194	0.131	0.163
5	0.250	0.229	0.239	0.362	0.362	0.362	0.244	0.228	0.236	0.143	0.096	0.119
6	0.218	0.201	0.210	0.331	0.332	0.332	0.208	0.195	0.202	0.113	0.076	0.095
7	0.199	0.184	0.192	0.313	0.312	0.312	0.190	0.178	0.184	0.094	0.062	0.078
8	0.189	0.175	0.182	0.302	0.301	0.301	0.182	0.171	0.176	0.083	0.054	0.069
9	0.180	0.168	0.174	0.293	0.293	0.293	0.172	0.162	0.167	0.076	0.048	0.062
10	0.175	0.163	0.169	0.287	0.286	0.286	0.167	0.159	0.163	0.072	0.044	0.058
Qwen2.5-VL-72B-Instruct												
4	0.647	0.622	0.635	0.900	0.900	0.900	0.747	0.733	0.740	0.294	0.233	0.264
5	0.635	0.614	0.625	0.900	0.900	0.900	0.742	0.731	0.736	0.263	0.212	0.237
6	0.637	0.615	0.626	0.900	0.900	0.900	0.754	0.742	0.748	0.256	0.205	0.230
7	0.634	0.614	0.624	0.901	0.900	0.901	0.754	0.742	0.748	0.249	0.201	0.225
8	0.635	0.615	0.625	0.903	0.903	0.903	0.756	0.744	0.750	0.247	0.199	0.223
9	0.636	0.616	0.626	0.904	0.904	0.904	0.758	0.745	0.752	0.246	0.198	0.222
10	0.639	0.618	0.629	0.906	0.906	0.906	0.762	0.749	0.756	0.250	0.200	0.225
Qwen2.5-VL-7B-Instruct												
4	0.610	0.563	0.587	0.919	0.906	0.913	0.625	0.594	0.609	0.286	0.189	0.237
5	0.560	0.527	0.544	0.878	0.869	0.873	0.594	0.572	0.583	0.208	0.140	0.174
6	0.533	0.506	0.520	0.852	0.847	0.849	0.570	0.550	0.560	0.178	0.122	0.150
7	0.516	0.495	0.506	0.838	0.837	0.837	0.557	0.540	0.548	0.154	0.108	0.131
8	0.506	0.488	0.497	0.832	0.832	0.832	0.548	0.532	0.540	0.139	0.099	0.119
9	0.500	0.483	0.491	0.825	0.829	0.827	0.544	0.528	0.536	0.133	0.093	0.113
10	0.497	0.480	0.488	0.823	0.826	0.825	0.542	0.526	0.534	0.126	0.088	0.107
Qwen3.5-Plus-2026-04-20												
4	0.680	0.661	0.671	0.933	0.939	0.936	0.758	0.744	0.751	0.347	0.300	0.324
5	0.651	0.632	0.641	0.918	0.925	0.921	0.714	0.697	0.706	0.322	0.274	0.298
6	0.639	0.622	0.631	0.902	0.914	0.908	0.696	0.682	0.689	0.319	0.269	0.294
7	0.631	0.614	0.623	0.897	0.906	0.901	0.687	0.673	0.680	0.309	0.262	0.285
8	0.628	0.611	0.619	0.893	0.906	0.899	0.687	0.672	0.679	0.303	0.256	0.279
9	0.625	0.610	0.617	0.890	0.903	0.897	0.684	0.669	0.677	0.302	0.257	0.279
10	0.624	0.611	0.617	0.888	0.905	0.897	0.683	0.667	0.675	0.300	0.260	0.280
Qwen3.6-Plus-2026-04-02												
4	0.697	0.683	0.690	0.942	0.947	0.944	0.783	0.786	0.784	0.367	0.317	0.342
5	0.670	0.657	0.663	0.924	0.933	0.929	0.749	0.747	0.748	0.338	0.292	0.315
6	0.657	0.645	0.651	0.909	0.919	0.914	0.733	0.733	0.733	0.330	0.281	0.305
7	0.646	0.636	0.641	0.903	0.916	0.909	0.718	0.720	0.719	0.317	0.273	0.295
8	0.640	0.633	0.637	0.896	0.914	0.905	0.715	0.715	0.715	0.309	0.269	0.289
9	0.638	0.631	0.635	0.894	0.914	0.904	0.714	0.711	0.712	0.307	0.268	0.287
10	0.636	0.630	0.633	0.889	0.912	0.901	0.712	0.708	0.710	0.306	0.270	0.288
Qwen-VL-Max												
4	0.707	0.699	0.703	0.967	0.972	0.970	0.831	0.819	0.825	0.325	0.306	0.316
5	0.697	0.691	0.694	0.963	0.969	0.966	0.812	0.808	0.810	0.317	0.296	0.306
6	0.694	0.690	0.692	0.958	0.967	0.962	0.809	0.806	0.808	0.316	0.295	0.305
7	0.692	0.686	0.689	0.958	0.965	0.962	0.803	0.799	0.801	0.315	0.293	0.304
8	0.690	0.685	0.688	0.957	0.964	0.960	0.803	0.798	0.800	0.312	0.293	0.302
9	0.690	0.685	0.688	0.956	0.964	0.960	0.800	0.796	0.798	0.312	0.296	0.304
10	0.689	0.687	0.688	0.956	0.963	0.960	0.800	0.794	0.797	0.313	0.302	0.307

Table 5: Full cumulative turn-indexed CorrTP@K results for all evaluated models on false-premise turns. O, Id, Attr, and Loc denote Overall, Identity, Attribute, and Location, respectively. GPT denotes the GPT-5.4 detector, Gemini denotes the Gemini-3.1-Pro-Preview detector, and Avg denotes the arithmetic mean of the two detector labels. We report only turns $K = 4$ to $K = 10$, because these are the false-premise turns in FPCO-Dialog.

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
Gemini 2.5 Flash												
4	0.650	0.641	0.645	0.928	0.944	0.936	0.756	0.753	0.754	0.267	0.225	0.246
5	0.672	0.649	0.661	0.964	0.967	0.966	0.781	0.769	0.775	0.272	0.211	0.241
6	0.659	0.648	0.653	0.964	0.967	0.966	0.772	0.769	0.770	0.242	0.208	0.225
7	0.651	0.638	0.645	0.956	0.956	0.956	0.767	0.761	0.764	0.231	0.197	0.214
8	0.668	0.652	0.660	0.972	0.964	0.968	0.783	0.775	0.779	0.247	0.217	0.232
9	0.662	0.648	0.655	0.958	0.956	0.957	0.781	0.772	0.776	0.247	0.217	0.232
10	0.665	0.653	0.659	0.961	0.964	0.962	0.786	0.775	0.780	0.247	0.219	0.233
Gemini 2.5 Pro												
4	0.704	0.688	0.696	0.947	0.947	0.947	0.811	0.800	0.806	0.353	0.317	0.335
5	0.657	0.651	0.654	0.931	0.936	0.933	0.742	0.733	0.738	0.300	0.283	0.291
6	0.656	0.646	0.651	0.931	0.931	0.931	0.744	0.744	0.744	0.294	0.264	0.279
7	0.630	0.623	0.627	0.917	0.925	0.921	0.706	0.703	0.704	0.267	0.242	0.255
8	0.638	0.640	0.639	0.917	0.922	0.919	0.717	0.728	0.722	0.281	0.269	0.275
9	0.644	0.636	0.640	0.919	0.922	0.921	0.725	0.731	0.728	0.286	0.256	0.271
10	0.640	0.638	0.639	0.917	0.922	0.919	0.725	0.714	0.720	0.278	0.278	0.278
Gemini 3.1 Pro Preview												
4	0.606	0.601	0.603	0.808	0.822	0.815	0.647	0.658	0.653	0.361	0.322	0.342
5	0.590	0.582	0.586	0.806	0.811	0.808	0.639	0.647	0.643	0.325	0.289	0.307
6	0.592	0.586	0.589	0.792	0.800	0.796	0.647	0.650	0.649	0.336	0.308	0.322
7	0.578	0.574	0.576	0.797	0.800	0.798	0.631	0.639	0.635	0.306	0.283	0.294
8	0.588	0.582	0.585	0.797	0.803	0.800	0.647	0.650	0.649	0.319	0.294	0.306
9	0.591	0.584	0.587	0.800	0.811	0.806	0.647	0.647	0.647	0.325	0.294	0.309
10	0.591	0.588	0.589	0.808	0.811	0.810	0.642	0.647	0.645	0.322	0.306	0.314
GPT-4.1												
4	0.619	0.611	0.615	0.917	0.917	0.917	0.689	0.678	0.683	0.253	0.239	0.246
5	0.601	0.587	0.594	0.911	0.908	0.909	0.678	0.669	0.673	0.214	0.183	0.199
6	0.592	0.585	0.589	0.903	0.908	0.905	0.661	0.661	0.661	0.211	0.186	0.199
7	0.583	0.579	0.581	0.897	0.903	0.900	0.653	0.642	0.647	0.200	0.192	0.196
8	0.581	0.569	0.575	0.892	0.900	0.896	0.650	0.636	0.643	0.200	0.172	0.186
9	0.582	0.573	0.577	0.897	0.900	0.899	0.647	0.636	0.641	0.203	0.183	0.193
10	0.587	0.573	0.580	0.889	0.903	0.896	0.658	0.633	0.645	0.214	0.183	0.199
GPT-4o												
4	0.523	0.490	0.506	0.878	0.872	0.875	0.489	0.467	0.478	0.203	0.131	0.167
5	0.500	0.461	0.481	0.850	0.853	0.851	0.467	0.456	0.462	0.183	0.075	0.129
6	0.501	0.461	0.481	0.853	0.850	0.851	0.467	0.456	0.462	0.183	0.078	0.131
7	0.495	0.456	0.476	0.850	0.850	0.850	0.469	0.453	0.461	0.167	0.064	0.116
8	0.499	0.457	0.478	0.853	0.850	0.851	0.467	0.453	0.460	0.178	0.069	0.123
9	0.499	0.456	0.478	0.858	0.856	0.857	0.467	0.450	0.459	0.172	0.061	0.116
10	0.494	0.456	0.475	0.842	0.844	0.843	0.472	0.461	0.467	0.169	0.061	0.115
GPT-5												
4	0.072	0.075	0.073	0.125	0.133	0.129	0.069	0.083	0.076	0.022	0.008	0.015
5	0.069	0.077	0.073	0.117	0.128	0.122	0.083	0.097	0.090	0.008	0.006	0.007
6	0.075	0.077	0.076	0.136	0.133	0.135	0.075	0.089	0.082	0.014	0.008	0.011
7	0.081	0.086	0.083	0.144	0.150	0.147	0.089	0.103	0.096	0.008	0.006	0.007
8	0.080	0.081	0.081	0.139	0.142	0.141	0.089	0.100	0.095	0.011	0.003	0.007
9	0.081	0.085	0.083	0.144	0.147	0.145	0.092	0.108	0.100	0.006	0.000	0.003
10	0.081	0.084	0.083	0.150	0.153	0.151	0.083	0.097	0.090	0.008	0.003	0.005
GPT-5.4												
4	0.376	0.360	0.368	0.575	0.542	0.558	0.433	0.425	0.429	0.119	0.114	0.116
5	0.326	0.318	0.322	0.492	0.475	0.483	0.389	0.383	0.386	0.097	0.094	0.096
6	0.295	0.281	0.288	0.406	0.392	0.399	0.369	0.364	0.366	0.111	0.086	0.099
7	0.280	0.269	0.275	0.375	0.361	0.368	0.367	0.364	0.365	0.097	0.083	0.090
8	0.277	0.267	0.272	0.383	0.369	0.376	0.364	0.361	0.362	0.083	0.069	0.076
9	0.269	0.263	0.266	0.358	0.358	0.358	0.367	0.356	0.361	0.083	0.075	0.079
10	0.271	0.263	0.267	0.364	0.364	0.364	0.353	0.353	0.353	0.097	0.072	0.084
InternVL2.5-2B												
4	0.139	0.131	0.135	0.178	0.175	0.176	0.094	0.092	0.093	0.144	0.128	0.136
5	0.058	0.054	0.056	0.097	0.086	0.091	0.022	0.025	0.024	0.056	0.050	0.053
6	0.040	0.037	0.038	0.069	0.067	0.068	0.019	0.022	0.020	0.031	0.022	0.026
7	0.039	0.032	0.036	0.053	0.050	0.052	0.042	0.033	0.038	0.022	0.014	0.018
8	0.036	0.031	0.034	0.058	0.053	0.056	0.019	0.019	0.019	0.031	0.019	0.025

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
9	0.034	0.029	0.032	0.053	0.050	0.052	0.031	0.025	0.028	0.019	0.011	0.015
10	0.031	0.029	0.030	0.044	0.042	0.043	0.017	0.022	0.019	0.031	0.022	0.026
InternVL2.5-4B												
4	0.145	0.134	0.140	0.217	0.211	0.214	0.106	0.103	0.104	0.114	0.089	0.102
5	0.075	0.070	0.073	0.153	0.147	0.150	0.042	0.036	0.039	0.031	0.028	0.029
6	0.067	0.065	0.066	0.139	0.142	0.141	0.036	0.036	0.036	0.025	0.017	0.021
7	0.059	0.059	0.059	0.131	0.133	0.132	0.033	0.033	0.033	0.014	0.011	0.013
8	0.057	0.056	0.057	0.125	0.128	0.127	0.031	0.025	0.028	0.017	0.014	0.015
9	0.066	0.064	0.065	0.136	0.136	0.136	0.036	0.036	0.036	0.025	0.019	0.022
10	0.064	0.061	0.062	0.139	0.136	0.138	0.033	0.031	0.032	0.019	0.017	0.018
InternVL2.5-8B												
4	0.152	0.147	0.149	0.283	0.281	0.282	0.103	0.100	0.102	0.069	0.061	0.065
5	0.111	0.106	0.108	0.217	0.219	0.218	0.075	0.072	0.073	0.042	0.025	0.034
6	0.105	0.101	0.103	0.222	0.219	0.221	0.069	0.067	0.068	0.022	0.017	0.019
7	0.101	0.098	0.100	0.214	0.214	0.214	0.072	0.069	0.071	0.017	0.011	0.014
8	0.103	0.102	0.102	0.219	0.219	0.219	0.069	0.069	0.069	0.019	0.017	0.018
9	0.104	0.100	0.102	0.217	0.217	0.217	0.075	0.072	0.073	0.019	0.011	0.015
10	0.108	0.105	0.106	0.217	0.217	0.217	0.089	0.086	0.087	0.019	0.011	0.015
LLaVA-NeXT-13B												
4	0.165	0.148	0.157	0.292	0.289	0.290	0.094	0.083	0.088	0.108	0.072	0.090
5	0.115	0.092	0.104	0.203	0.194	0.199	0.053	0.047	0.050	0.089	0.033	0.061
6	0.093	0.078	0.085	0.156	0.158	0.157	0.047	0.036	0.041	0.075	0.039	0.057
7	0.079	0.074	0.076	0.150	0.153	0.151	0.050	0.044	0.047	0.036	0.025	0.030
8	0.083	0.071	0.077	0.150	0.147	0.148	0.050	0.039	0.044	0.050	0.028	0.039
9	0.081	0.069	0.075	0.142	0.136	0.139	0.042	0.042	0.042	0.058	0.028	0.043
10	0.081	0.067	0.074	0.144	0.142	0.143	0.039	0.031	0.035	0.058	0.028	0.043
LLaVA-OneVision-0.5B												
4	0.060	0.051	0.055	0.047	0.053	0.050	0.025	0.039	0.032	0.108	0.061	0.084
5	0.029	0.014	0.022	0.011	0.003	0.007	0.014	0.014	0.014	0.061	0.025	0.043
6	0.027	0.016	0.021	0.008	0.006	0.007	0.008	0.011	0.009	0.064	0.031	0.048
7	0.024	0.008	0.016	0.008	0.006	0.007	0.014	0.011	0.013	0.050	0.008	0.029
8	0.020	0.008	0.014	0.003	0.003	0.003	0.014	0.014	0.014	0.044	0.008	0.026
9	0.020	0.010	0.015	0.008	0.008	0.008	0.014	0.014	0.014	0.039	0.008	0.024
10	0.015	0.006	0.010	0.008	0.006	0.007	0.003	0.003	0.003	0.033	0.011	0.022
LLaVA-OneVision-7B												
4	0.021	0.031	0.026	0.019	0.042	0.030	0.025	0.039	0.032	0.019	0.014	0.017
5	0.011	0.010	0.010	0.014	0.008	0.011	0.011	0.014	0.013	0.008	0.008	0.008
6	0.008	0.008	0.008	0.003	0.006	0.005	0.014	0.017	0.015	0.008	0.003	0.005
7	0.006	0.005	0.005	0.003	0.003	0.003	0.006	0.008	0.007	0.008	0.003	0.005
8	0.011	0.005	0.008	0.003	0.003	0.003	0.008	0.008	0.008	0.022	0.003	0.012
9	0.006	0.006	0.006	0.003	0.006	0.005	0.008	0.011	0.009	0.008	0.003	0.005
10	0.005	0.004	0.005	0.006	0.003	0.005	0.003	0.008	0.005	0.006	0.000	0.003
Qwen2.5-VL-32B-Instruct												
4	0.629	0.620	0.625	0.894	0.900	0.897	0.717	0.700	0.708	0.275	0.261	0.268
5	0.573	0.574	0.573	0.847	0.853	0.850	0.661	0.667	0.664	0.211	0.203	0.207
6	0.569	0.576	0.573	0.864	0.869	0.867	0.653	0.667	0.660	0.192	0.192	0.192
7	0.556	0.568	0.562	0.847	0.858	0.853	0.642	0.653	0.647	0.178	0.192	0.185
8	0.553	0.559	0.556	0.858	0.869	0.863	0.647	0.642	0.645	0.153	0.167	0.160
9	0.554	0.550	0.552	0.850	0.844	0.847	0.636	0.642	0.639	0.175	0.164	0.169
10	0.554	0.554	0.554	0.856	0.872	0.864	0.650	0.631	0.641	0.156	0.158	0.157
Qwen2.5-VL-3B-Instruct												
4	0.316	0.283	0.299	0.425	0.417	0.421	0.328	0.303	0.316	0.194	0.131	0.163
5	0.184	0.174	0.179	0.300	0.308	0.304	0.161	0.153	0.157	0.092	0.061	0.076
6	0.153	0.146	0.149	0.269	0.272	0.271	0.136	0.131	0.134	0.053	0.036	0.044
7	0.144	0.134	0.139	0.258	0.253	0.256	0.136	0.128	0.132	0.036	0.022	0.029
8	0.147	0.138	0.143	0.256	0.256	0.256	0.147	0.139	0.143	0.039	0.019	0.029
9	0.138	0.130	0.134	0.247	0.250	0.248	0.122	0.119	0.120	0.044	0.019	0.032
10	0.144	0.136	0.140	0.250	0.244	0.247	0.139	0.142	0.141	0.044	0.022	0.033
Qwen2.5-VL-72B-Instruct												
4	0.647	0.622	0.635	0.900	0.900	0.900	0.747	0.733	0.740	0.294	0.233	0.264
5	0.622	0.606	0.614	0.900	0.900	0.900	0.736	0.728	0.732	0.231	0.192	0.212

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
6	0.641	0.618	0.629	0.900	0.900	0.900	0.778	0.764	0.771	0.244	0.189	0.216
7	0.628	0.611	0.619	0.903	0.900	0.901	0.756	0.744	0.750	0.225	0.189	0.207
8	0.639	0.619	0.629	0.914	0.914	0.914	0.761	0.750	0.756	0.242	0.192	0.217
9	0.641	0.619	0.630	0.908	0.911	0.909	0.772	0.753	0.762	0.242	0.192	0.217
10	0.656	0.634	0.645	0.919	0.917	0.918	0.781	0.769	0.775	0.269	0.217	0.243
Qwen2.5-VL-7B-Instruct												
4	0.610	0.563	0.587	0.919	0.906	0.913	0.625	0.594	0.609	0.286	0.189	0.237
5	0.510	0.492	0.501	0.836	0.833	0.835	0.564	0.550	0.557	0.131	0.092	0.112
6	0.480	0.465	0.473	0.800	0.803	0.802	0.522	0.506	0.514	0.117	0.086	0.102
7	0.465	0.459	0.462	0.794	0.806	0.800	0.517	0.508	0.512	0.083	0.064	0.074
8	0.467	0.459	0.463	0.811	0.814	0.812	0.511	0.500	0.506	0.078	0.064	0.071
9	0.470	0.461	0.466	0.786	0.811	0.798	0.522	0.508	0.515	0.103	0.064	0.083
10	0.477	0.461	0.469	0.811	0.811	0.811	0.533	0.514	0.524	0.086	0.058	0.072
Qwen3.5-Plus-2026-04-20												
4	0.680	0.661	0.671	0.933	0.939	0.936	0.758	0.744	0.751	0.347	0.300	0.324
5	0.623	0.603	0.613	0.903	0.911	0.907	0.669	0.650	0.659	0.297	0.247	0.272
6	0.614	0.601	0.607	0.869	0.892	0.881	0.661	0.653	0.657	0.311	0.258	0.284
7	0.606	0.590	0.598	0.881	0.883	0.882	0.658	0.644	0.651	0.281	0.242	0.262
8	0.615	0.601	0.608	0.881	0.903	0.892	0.686	0.667	0.677	0.278	0.233	0.256
9	0.613	0.604	0.609	0.872	0.892	0.882	0.669	0.656	0.663	0.297	0.264	0.280
10	0.616	0.616	0.616	0.881	0.914	0.897	0.675	0.656	0.665	0.292	0.278	0.285
Qwen3.6-Plus-2026-04-02												
4	0.697	0.683	0.690	0.942	0.947	0.944	0.783	0.786	0.784	0.367	0.317	0.342
5	0.643	0.631	0.637	0.906	0.919	0.913	0.714	0.708	0.711	0.308	0.267	0.287
6	0.632	0.619	0.625	0.881	0.892	0.887	0.703	0.706	0.704	0.314	0.261	0.287
7	0.611	0.611	0.611	0.883	0.906	0.895	0.672	0.681	0.677	0.278	0.247	0.263
8	0.618	0.619	0.619	0.869	0.908	0.889	0.703	0.694	0.698	0.281	0.256	0.269
9	0.630	0.621	0.625	0.883	0.914	0.899	0.708	0.689	0.698	0.297	0.261	0.279
10	0.619	0.624	0.621	0.858	0.900	0.879	0.697	0.689	0.693	0.300	0.283	0.291
Qwen-VL-Max												
4	0.707	0.699	0.703	0.967	0.972	0.970	0.831	0.819	0.825	0.325	0.306	0.316
5	0.687	0.683	0.685	0.958	0.967	0.962	0.794	0.797	0.796	0.308	0.286	0.297
6	0.689	0.686	0.688	0.950	0.961	0.956	0.803	0.803	0.803	0.314	0.294	0.304
7	0.683	0.674	0.679	0.956	0.961	0.958	0.783	0.775	0.779	0.311	0.286	0.298
8	0.685	0.683	0.684	0.953	0.961	0.957	0.803	0.797	0.800	0.300	0.292	0.296
9	0.685	0.686	0.685	0.953	0.961	0.957	0.789	0.783	0.786	0.314	0.314	0.314
10	0.689	0.694	0.692	0.956	0.961	0.958	0.794	0.786	0.790	0.317	0.333	0.325

Table 6: Full turn-wise TurnCorr@K results for all evaluated models on false-premise turns. O, Id, Attr, and Loc denote Overall, Identity, Attribute, and Location, respectively. GPT denotes the GPT-5.4 detector, Gemini denotes the Gemini-3.1-Pro-Preview detector, and Avg denotes the arithmetic mean of the two detector labels. We report only turns $K = 4$ to $K = 10$, because these are the false-premise turns in FPCO-Dialog.

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
Gemini 2.5 Flash												
1	0.020	0.018	0.019	0.017	0.011	0.014	0.036	0.033	0.035	0.008	0.008	0.008
2	0.020	0.017	0.019	0.015	0.011	0.013	0.033	0.032	0.033	0.011	0.008	0.009
3	0.019	0.017	0.018	0.014	0.012	0.013	0.033	0.031	0.032	0.009	0.007	0.008
Gemini 2.5 Pro												
1	0.026	0.019	0.022	0.017	0.011	0.014	0.039	0.031	0.035	0.022	0.014	0.018
2	0.023	0.015	0.019	0.017	0.014	0.015	0.032	0.024	0.028	0.019	0.008	0.013
3	0.020	0.013	0.017	0.016	0.011	0.013	0.028	0.020	0.024	0.017	0.006	0.011
Gemini 3.1 Pro Preview												
1	0.031	0.026	0.028	0.025	0.022	0.024	0.042	0.033	0.038	0.025	0.022	0.024
2	0.025	0.021	0.023	0.018	0.017	0.018	0.036	0.029	0.033	0.021	0.017	0.019
3	0.022	0.019	0.020	0.017	0.015	0.016	0.032	0.028	0.030	0.018	0.014	0.016
GPT-4.1												
1	0.012	0.009	0.010	0.003	0.003	0.003	0.019	0.011	0.015	0.014	0.014	0.014
2	0.011	0.008	0.009	0.003	0.003	0.003	0.015	0.008	0.011	0.015	0.014	0.014

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
3	0.010	0.008	0.009	0.003	0.003	0.003	0.013	0.007	0.010	0.014	0.013	0.013
GPT-4o												
1	0.010	0.008	0.009	0.008	0.006	0.007	0.017	0.014	0.015	0.006	0.006	0.006
2	0.009	0.008	0.009	0.007	0.006	0.007	0.015	0.015	0.015	0.006	0.003	0.005
3	0.008	0.007	0.007	0.006	0.005	0.005	0.015	0.014	0.014	0.005	0.002	0.004
GPT-5												
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
GPT-5.4												
1	0.001	0.000	0.001	0.000	0.000	0.000	0.003	0.000	0.002	0.000	0.000	0.000
2	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.001	0.000	0.001
3	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.001	0.000	0.001
InternVL2.5-2B												
1	0.015	0.009	0.012	0.006	0.006	0.006	0.025	0.014	0.019	0.014	0.008	0.011
2	0.010	0.007	0.009	0.007	0.006	0.007	0.015	0.011	0.013	0.008	0.004	0.006
3	0.009	0.005	0.007	0.005	0.004	0.005	0.014	0.009	0.011	0.009	0.003	0.006
InternVL2.5-4B												
1	0.005	0.004	0.005	0.006	0.003	0.005	0.006	0.006	0.006	0.003	0.003	0.003
2	0.004	0.003	0.004	0.004	0.003	0.004	0.007	0.006	0.007	0.001	0.001	0.001
3	0.003	0.003	0.003	0.004	0.003	0.004	0.006	0.005	0.005	0.001	0.001	0.001
InternVL2.5-8B												
1	0.006	0.003	0.005	0.006	0.000	0.003	0.011	0.006	0.009	0.003	0.003	0.003
2	0.005	0.002	0.004	0.003	0.000	0.002	0.008	0.006	0.007	0.003	0.001	0.002
3	0.005	0.002	0.004	0.002	0.000	0.001	0.007	0.006	0.007	0.005	0.001	0.003
LLaVA-NeXT-13B												
1	0.006	0.005	0.005	0.003	0.003	0.003	0.008	0.006	0.007	0.006	0.006	0.006
2	0.003	0.003	0.003	0.001	0.001	0.001	0.006	0.004	0.005	0.003	0.003	0.003
3	0.002	0.002	0.002	0.001	0.001	0.001	0.005	0.004	0.005	0.002	0.002	0.002
LLaVA-OneVision-0.5B												
1	0.004	0.001	0.003	0.003	0.000	0.002	0.006	0.003	0.005	0.003	0.000	0.002
2	0.003	0.000	0.002	0.003	0.000	0.002	0.004	0.001	0.003	0.001	0.000	0.001
3	0.002	0.000	0.001	0.003	0.000	0.002	0.003	0.001	0.002	0.002	0.000	0.001
LLaVA-OneVision-7B												
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Qwen2.5-VL-32B-Instruct												
1	0.033	0.028	0.030	0.031	0.028	0.029	0.031	0.028	0.029	0.039	0.028	0.034
2	0.029	0.024	0.027	0.022	0.019	0.020	0.032	0.028	0.030	0.033	0.025	0.029
3	0.029	0.024	0.027	0.021	0.021	0.021	0.031	0.029	0.030	0.033	0.023	0.028
Qwen2.5-VL-3B-Instruct												
1	0.006	0.002	0.004	0.006	0.003	0.005	0.006	0.003	0.005	0.006	0.000	0.003
2	0.003	0.001	0.002	0.003	0.001	0.002	0.003	0.001	0.002	0.003	0.000	0.002
3	0.003	0.001	0.002	0.003	0.001	0.002	0.003	0.001	0.002	0.003	0.000	0.002
Qwen2.5-VL-72B-Instruct												
1	0.028	0.024	0.026	0.025	0.022	0.024	0.036	0.031	0.034	0.022	0.019	0.020
2	0.024	0.022	0.023	0.018	0.017	0.018	0.033	0.031	0.032	0.021	0.019	0.020
3	0.023	0.023	0.023	0.018	0.017	0.018	0.032	0.031	0.032	0.020	0.019	0.019
Qwen2.5-VL-7B-Instruct												
1	0.020	0.016	0.018	0.025	0.019	0.022	0.025	0.019	0.022	0.011	0.008	0.009
2	0.018	0.015	0.017	0.018	0.014	0.016	0.025	0.021	0.023	0.011	0.010	0.010
3	0.017	0.014	0.015	0.016	0.012	0.014	0.025	0.021	0.023	0.010	0.009	0.009
Qwen3.5-Plus-2026-04-20												
1	0.024	0.020	0.022	0.014	0.014	0.014	0.036	0.025	0.030	0.022	0.022	0.022
2	0.021	0.018	0.019	0.014	0.014	0.014	0.029	0.021	0.025	0.019	0.018	0.018
3	0.016	0.014	0.015	0.010	0.010	0.010	0.023	0.018	0.020	0.016	0.014	0.015

K	O-GPT	O-Gem	O-Avg	Id-GPT	Id-Gem	Id-Avg	Attr-GPT	Attr-Gem	Attr-Avg	Loc-GPT	Loc-Gem	Loc-Avg
						Qwen3.6-Plus-2026-04-02						
1	0.017	0.017	0.017	0.006	0.011	0.009	0.025	0.028	0.027	0.019	0.011	0.015
2	0.016	0.016	0.016	0.010	0.011	0.010	0.025	0.028	0.027	0.014	0.008	0.011
3	0.015	0.015	0.015	0.009	0.010	0.009	0.024	0.024	0.024	0.012	0.009	0.010
						Qwen-VL-Max						
1	0.017	0.016	0.017	0.014	0.014	0.014	0.019	0.019	0.019	0.017	0.014	0.015
2	0.017	0.017	0.017	0.013	0.014	0.013	0.019	0.019	0.019	0.018	0.017	0.018
3	0.017	0.016	0.017	0.013	0.014	0.013	0.019	0.019	0.019	0.019	0.016	0.018

Table 7: Full cumulative CorrFP@K results for all evaluated models on premise-correct turns. O denotes Overall, while Id, Attr, and Loc denote dialogues assigned to the Identity, Attribute, and Location false-premise classes, respectively. GPT denotes the GPT-5.4 detector, Gemini denotes the Gemini-3.1-Pro-Preview detector, and Avg denotes the arithmetic mean of the two detector labels. We report turns $K = 1$ to $K = 3$, because these are the premise-correct turns in FPCO-Dialog.