DIXITWORLD: Evaluating Multimodal Abductive Reasoning in Vision-Language Models with Multi-Agent Dixit Gameplay

Yunxiang Mo*, **Tianshi Zheng***, **Qing Zong, Jiayu Liu, Baixuan Xu**
Yauwai Yim, Chunkit Chan, Jiaxin Bai, Yangqiu Song
Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
{ymoaj, tzhengad}@connect.ust.hk, yqsong@cse.ust.hk

Abstract

Multimodal abductive reasoning—the generation and selection of explanatory hypotheses from partial observations—is a cornerstone of intelligence. Current evaluations of such ability in vision–language models (VLMs) are largely confined to static, single-agent tasks. Inspired by *Dixit*, we introduce DIXITWORLD[1], a comprehensive evaluation suite designed to deconstruct this challenge. DIXITWORLD features two core components: **DixitArena**, a dynamic, multi-agent environment that evaluates both hypothesis generation (a "storyteller" crafting cryptic clues) and hypothesis selection ("listeners" choosing the target image from decoys) under imperfect information; and **DixitBench**, a static QA benchmark that isolates the listener's task for efficient, controlled evaluation. Results from DixitArena reveal distinct, role-dependent behaviors: smaller open-source models often excel as creative storytellers, producing imaginative yet less discriminative clues, whereas larger proprietary models demonstrate superior overall performance, particularly as listeners. Performance on DixitBench strongly correlates with listener results in DixitArena, validating it as a reliable proxy for hypothesis selection. Our findings reveal a key trade-off between generative creativity and discriminative understanding in multimodal abductive reasoning, a central challenge for developing more balanced and capable vision-language agents.

1 Introduction

Abductive reasoning (Peirce, 1931; Frankfurt, 1958) is a cornerstone of human intelligence denoting the inference process towards the best explanatory hypothesis from observation. This ability is fundamental to complex cognitive tasks ranging from commonsense reasoning to scientific discovery (Bhagavatula et al., 2020; Bisk et al., 2019).

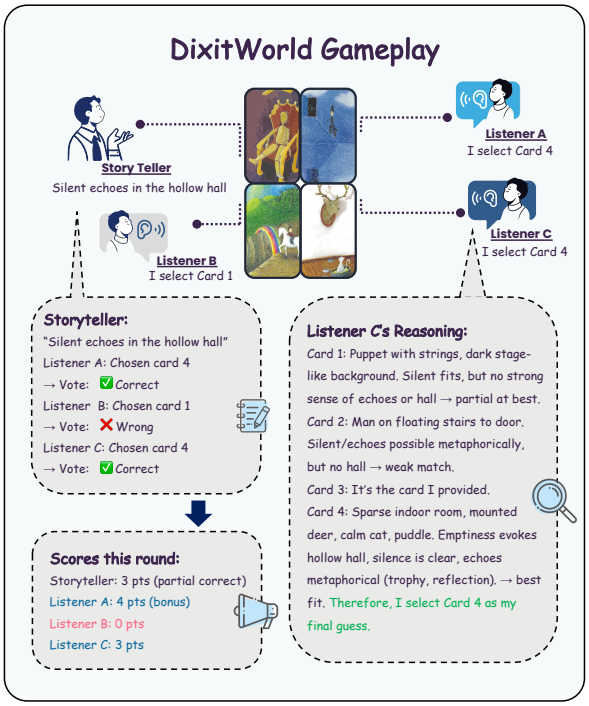

Figure 1: An illustration of Dixit gameplay.

Recent advances in vision-language models (VLMs) (Liu et al., 2023; Google, 2025) have shown remarkable progress in perception-grounded reasoning tasks. However, the extent to which these models can perform abductive reasoning remains unclear. Prior benchmarks such as VAR (Liang et al., 2022) and Sherlock (Hessel et al., 2022) are limited to static, single-agent inference with fixed candidate sets, leaving open the question of how VLMs behave when abductive reasoning must be both generated and evaluated in dynamic, multi-agent scenarios.

To bridge this gap, we introduce DIXITWORLD, an evaluation suite inspired by the game *Dixit*. Its core component, **DixitArena**, provides a near-perfect operationalization of multimodal abductive reasoning by naturally decomposing the process into two complementary roles. The **Storyteller** performs **hypothesis generation**: given an image

* Equal Contribution
[1] https://github.com/HKUST-KnowComp/DixitWorld

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 448–462
July 2-7, 2026

(the observation), they must generate a cryptic clue (the explanatory hypothesis) that is abstract enough to create ambiguity but clear enough to be solvable. Conversely, the **Listeners** perform **hypothesis selection**: presented with the clue, they must perform an "inference to the best explanation" by evaluating a set of competing visual hypotheses—the target and adversarial decoys—to identify the most plausible origin of the clue. Our large-scale simulations with this setup reveal a stark asymmetry in how models handle these roles: smaller open-source models often act as more creative but less precise storytellers, while larger proprietary models excel as listeners but struggle with the storyteller's core challenge of balancing informativeness and ambiguity.

To enable more efficient and controlled evaluation, we further curate **DixitBench**. This benchmark isolates the listener's role, reframing the hypothesis selection challenge as a multiple-choice QA task with adjustable difficulty. Crucially, model performance on DixitBench strongly correlates with listener performance in DixitArena, validating it as a reliable and lightweight proxy for evaluating hypothesis selection.

Taken together, DIXITWORLD offers a threefold contribution: (1) we introduce DixitArena and DixitBench as complementary benchmarks that decompose multimodal abductive reasoning into generation and selection under adversarial conditions; (2) we reveal a critical storyteller–listener asymmetry—over 78% of storyteller rounds yield zero points, yet listener accuracy reaches 75.6% for the best model—demonstrating that discrimination far outpaces controlled creative generation in current VLMs; and (3) we provide extensive analyses (scaling, calibration, sensitivity) showing that this gap is structural, not attributable to model scale or evaluation artifacts ($r = 0.947$ between DixitBench and Arena).

2 DIXITWORLD

DixitArena DixitArena is an interactive, multi-agent environment designed to operationalize multimodal abductive reasoning. In each match, four agents alternate between the roles of **Storyteller** and **Listeners**. The Storyteller performs hypothesis generation by observing an image and creating a cryptic clue. The Listeners then perform hypothesis selection, inferring which image best explains that clue from a set of distractors. A key feature is the scoring mechanism, which directly rewards abductive success: the Storyteller only scores points for partial correctness—when some, but not all, Listeners identify the target.

DixitBench To complement the dynamic environment, DixitBench is a static benchmark designed to isolate and control the evaluation of the listener's task. It reframes hypothesis selection as a multiple-choice QA problem, where models must identify the target image for a given clue from a set of distractors. The benchmark consists of 252 questions (84 images $\times$ 3 difficulty tiers: Easy, Medium, Hard), where each item has 1 target and 5 distractors (6 candidates total) and distractor difficulty is systematically controlled via semantic similarity between pre-generated captions. In the main text we focus on the *Easy* and *Hard* tiers (168 items) for a clear difficulty contrast; the *Medium* tier is reported in Appendix E. The framework is highly scalable: with 84 images, the total possible question configurations (1 target + 5 distractors) number $\binom{84}{6} \approx 4.06 \times 10^8$. The current 252 items constitute a balanced, analyzable subset validated by human annotation ($\kappa > 0.8$; see Appendix E). As performance on DixitBench strongly correlates with listener results in DixitArena, it serves as a validated and efficient proxy for this ability.

Evaluation Metrics Performance in **DixitArena** is measured using three primary metrics: (i) **Storyteller Score**, the percentage of rounds achieving partial correctness; (ii) **Listener Accuracy**, the percentage of correct target identifications; and (iii) **Overall Score**, the normalized total points earned per match (see Eq. 1), reflecting the combined success across both roles under the Dixit scoring rules.

The overall normalized score for a model is the average of its normalized scores across all matches, calculated as:

$$Score = \left(\frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \frac{\text{score}_i^{\text{attained}}}{\text{score}_i^{\text{max}}} \right) \times 100\% \quad (1)$$

where $\mathcal{N}$ represents the set of matches played.

For DixitBench, we report overall accuracy as well as performance across the Easy and Hard subsets. All scores are normalized to a 0–100% scale. Full implementation details, fairness procedures, and difficulty calibration are provided in Appendix E and D.

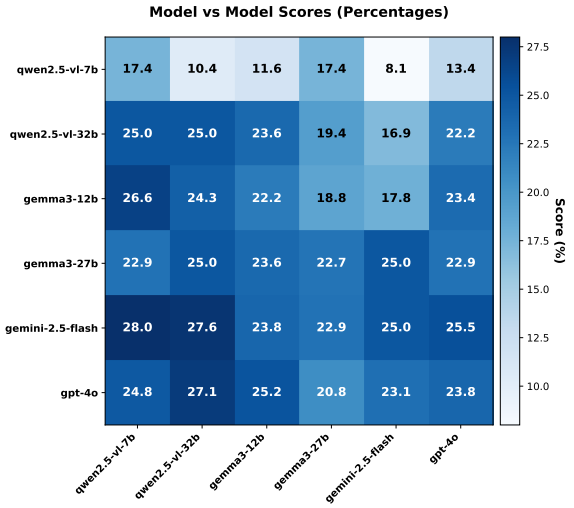

Figure 2: Head-to-head normalized overall scores (%) in the round-robin tournament. Each cell shows the row model's score when matched against the column model. Higher values indicate stronger performance against that opponent.

Model	Storyteller	Listener	Overall
Qwen2.5-VL-7B	14.29	30.87	17.39
Qwen2.5-VL-32B	15.48	50.79	30.63
Gemma3-12B	20.24	49.44	31.68
Gemma3-27B	**32.14**	49.76	34.15
GPT-4o	19.05	<u>53.89</u>	<u>34.58</u>
Gemini-2.5-Flash	<u>30.95</u>	**54.05**	**36.52**

Table 1: Normalized scores (%) for Storyteller, Listener, and Overall performance in DixitArena. Models are ranked by their overall average score. The best and second-best scores in each column are in **bold** and <u>underlined</u>, respectively.

3 Experiment and Analysis

We evaluate six modern VLMs of varying scales, including Qwen2.5-VL-7B/32B, Gemini-2.5-Flash, Gemma3-12B/27B, and GPT-4o. All models are configured with the temperature of 0.7 to encourage diversity. Model specifications and prompt templates are detailed in Appendix B and C.

3.1 DixitArena Gameplay Performance

Figure 2 illustrates the head-to-head gameplay performances in our round-robin tournament across all six models in DixitArena.

Larger and proprietary models (*Gemini-2.5-Flash*, *GPT-4o*) generally outperform smaller open-source ones, confirming that abductive reasoning benefits from scale and more comprehensive multimodal pretraining. In addition, models' scores are influenced by the strength of their opponents—when facing stronger models with higher

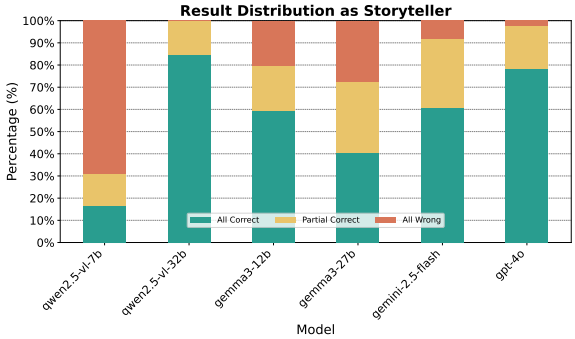

Figure 3: Distribution of storyteller round outcomes. Only the "Partial-Correct" outcome yields points for the storyteller, making it the desired result.

average performance, weaker models' scores tend to drop, indicating that high-performing agents impose stronger abductive pressure through clearer or more discriminative cues. Conversely, matches between weaker models exhibit higher variance and less stable gameplay performance. Overall, the matrix reveals a consistent capability hierarchy alongside non-trivial interaction effects between model strength and opponent style.

Furthermore, Table 1 presents the gameplay performance with different roles isolated. Listener performance generally scales with model size and proprietary status, while storyteller ability does not. This reveals that while large models are adept at *understanding* clues, they struggle with *crafting* them with the requisite balance of ambiguity and clarity. For instance, GPT-4o excels as a listener yet ranks only mid-level as a storyteller, underscoring this performance gap. A scaling analysis with an additional 72B-parameter model (Appendix I) confirms that this asymmetry persists at frontier open-source scale, ruling out insufficient model capacity as an explanation.

To further diagnose the asymmetry in storyteller performance, we decomposed round outcomes into three categories: *Partial-Correct* (the optimal outcome), *All-Correct* (too obvious), and *All-Wrong* (too vague or misleading). As shown in Figure 3, over 78% of storyteller rounds yielded zero points, falling into either the All-Correct or All-Wrong categories. This high failure rate reveals that models frequently struggle to strike the intended "Dixit balance" between clarity and ambiguity. The nature of this failure varies by model scale: smaller models often produce overly literal or vague clues, while larger ones tend toward over-specificity. This confirms that generative abduction—requiring a blend

Model	Easy	Hard	Total
Qwen2.5-VL-7B	39.05	34.29	36.67
Qwen2.5-VL-32B	70.24	55.95	63.10
Gemma3-12B	58.33	57.14	57.74
Gemma3-27B	63.10	61.90	62.50
Gemini-2.5-Flash	65.48	64.29	64.89
GPT-4o	**78.57**	**72.62**	**75.60**

Table 2: Performance on the DixitBench. The small gap between Easy and Hard subsets suggests limited sensitivity to semantic difficulty.

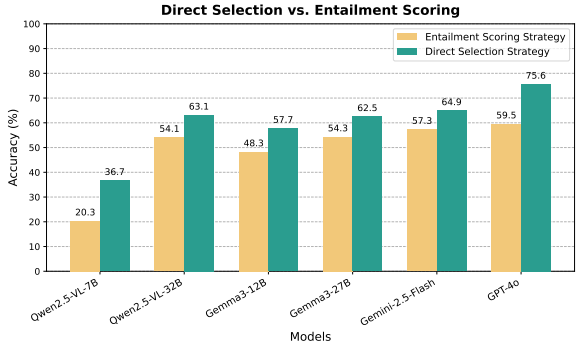

Figure 4: Comparison between direct selection and entailment scoring strategies in DixitBench.

of creativity and controlled uncertainty—remains a primary bottleneck for current VLMs.

3.2 DixitBench Performance

Table 2 presents VLM performances on DixitBench, which are highly consistent with listener results from DixitArena (Pearson $r = 0.947$, $p = 0.004$, 95% CI $[0.65, 0.99]$; Spearman $\rho = 0.929$, $p = 0.003$), thus validating the benchmark's reliability as a proxy for hypothesis selection. The results also show that the difficulty classification based on semantic distance yields only a modest performance gap (4.8% in average) between *Easy* and *Hard* questions. Although directionally valid, this modest gap shows that simple semantic distance is not a strong indicator of difficulty for the abstract gameplay scenarios in Dixit. Future iterations of DixitBench will explore factorial distractors (varying object, relation, style, and affect independently) to achieve larger difficulty separation.

We further compared listener performance under two hypothesis selection strategies: *direct selection* (choosing one image) and *entailment scoring* (rating all candidates independently and choose the highest one). As shown in Figure 4, all models perform consistently better under direct selection, with accuracy drops of roughly 5–20 percentage points when switching to entailment scoring (complete results in Appendix M.1). These results indicate that absolute plausibility rating introduces calibration noise and reduces model separability, particularly when distractors are semantically close. A calibration analysis (Appendix P) confirms this: all models exhibit substantial ECE values (> 0.17), with GPT-4o showing the best calibration (Brier $= 0.22$) and smaller models showing larger confidence-accuracy gaps. In short, VLMs are more reliable at *selective abductive inference* (choosing among alternatives) than at *discriminative entailment judgment*.

3.3 Human Evaluation

To complement our automated metrics, we conducted a human evaluation of the storyteller clues. We assessed the clues on two rated dimensions, **Clarity** and **Creativity**, and one performance-based dimension, **Listener Accuracy**. For the rated dimensions, human annotators scored each clue with an inter-rater agreement of Cohen's $\kappa \approx 0.75$. Clarity was then calculated using our three-step robust index (detailed in Appendix G), while creativity scores represent the normalized mean of the ratings. Listener Accuracy was measured directly as the success rate of human players using the AI-generated clues.

As demonstrated in Table 3, human listeners achieved 71.6% accuracy with the AI clues, a score closely matched by Gemini-2.5-Flash (70.2%). Among the models, a trade-off was evident: Gemma3-27B achieved the highest clarity balance (0.387), while Gemma3-12B scored highest on creativity (0.574). These results highlight a tension between informativeness and imagination, indicating that effective abductive communication demands *moderate clarity with nontrivial abstraction*. This mirrors the storyteller failure patterns observed earlier, where overly explicit models lose ambiguity and overly creative ones lose alignment with the target image.

3.4 Further Discussions

We performed several checks to validate our evaluation's robustness and fairness. A leave-one-listener-out (LOLO) analysis confirmed stable scores across listener subsets (Appendix M.2), while a chi-squared test showed a uniform distribution of candidate image positions ($\chi^2 = 1.83, p = 0.61$). Hand-swap phases also effectively minimized bias, with score differences remaining under 0.3 points. These results affirm the stability and impartiality

Model	Acc (%)	Clarity	Creativity
Qwen2.5-VL-7B	25.79	0.104	0.212
Qwen2.5-VL-32B	66.27	0.188	0.243
Gemma3-12B	63.49	0.295	**0.574**
Gemma3-27B	65.08	**0.387**	0.536
Gemini-2.5-Flash	70.24	0.310	0.488
GPT-4o	68.65	0.232	0.484
Human	**71.59**	—	—

Table 3: Human evaluation results. "Clarity" is computed using our three-step median–penalty formula. Creativity is normalized to [0,1].

of our environment.

Importantly, our design isolates abductive reasoning from potential caption artifacts through three mechanisms: (i) the Dixit scoring rule naturally penalizes both over-literal and misleading clues by awarding zero points unless partial correctness is achieved; (ii) human evaluation uses the true target image, so clarity and creativity ratings reflect pragmatic quality rather than caption artifacts; and (iii) captions are used solely for distractor sampling during benchmark construction—listeners receive only raw images at inference time, ensuring evaluation is based purely on visual semantics.

A per-image category analysis (Appendix L) reveals further nuance: GPT-4o shows a pronounced advantage on metaphorical images (+17.1pp over literal scenes), suggesting superior abstract reasoning, while Gemma3-27B exhibits the opposite pattern (-6.4pp), relying more on concrete visual features. This asymmetry in image-type sensitivity parallels the broader storyteller–listener gap, as metaphorical scenes demand the same kind of abstract reasoning required for effective clue generation. A seed sweep across three random seeds (Appendix Q) confirms evaluation stability, with all variances below ± 3.0pp.

Overall, our analysis indicates that while current VLMs possess strong discriminative grounding, they lack sophisticated pragmatic control. They succeed at hypothesis selection but fail to navigate the communicative ambiguity required for hypothesis generation. Bridging this storyteller–listener asymmetry will likely require explicitly modeling theory-of-mind, uncertainty, and ambiguity.

3.5 Error Analysis

Qualitative analysis of storyteller failures reveals two distinct error modes. *Over-specification* occurs when models produce literal descriptions (e.g., *"A knight in shining armor on horseback, holding a spear"*), leaving no room for ambiguity—all listeners guess correctly and the storyteller scores zero. Conversely, *semantic mismatch* arises from vague or misaligned clues (e.g., *"A forgotten trophy, watching over a curious secret"*) that fail to connect with the target image, causing all listeners to guess incorrectly. These opposing failure patterns (detailed in Appendix K) underscore that effective abductive generation requires navigating a narrow corridor between clarity and ambiguity—a capability that current VLMs, regardless of scale, have not yet acquired.

4 Conclusion

This paper demonstrates that current VLMs, despite their strong discriminative abilities, lack the pragmatic control necessary for robust multimodal abductive reasoning. Through our novel multi-agent benchmark, DixitArena, we reveal a critical performance gap: models are proficient listeners (hypothesis selectors) but poor storytellers (hypothesis generators), struggling to balance creativity with communicative intent. This core failure to manage ambiguity highlights that discriminative success in static QA tasks does not guarantee generative competence in dynamic, goal-oriented scenarios. Advancing the field will therefore require a shift towards architectures that explicitly reason about ambiguity and intention modeling.

Limitations

Our study has several limitations that future work may address:

- **Task Scope.** While we operationalize abductive reasoning through the game *Dixit*, it represents only one class of multimodal abductive settings. Real-world abductive reasoning often involves temporal, causal, or interactive elements that our static and turn-based setup does not fully capture. A natural extension is to short videos or ordered image panels, enabling evaluation of causal and temporal abductive reasoning.

- **Model Coverage.** We evaluated a representative but limited set of modern VLMs (Qwen, Gemma, Gemini, GPT families). Broader coverage—including models fine-tuned for visual entailment or narrative generation—would further validate our conclusions about role asymmetry and entailment calibration.

- **Prompt and Decoding Sensitivity.** We used a standardized prompting and decoding configuration (temperature 0.7, JSON output) for comparability. Performance may vary with prompt wording, sampling parameters, or instruction tuning. A systematic study of strategy sensitivity remains future work.

- **Evaluation Strategy Dependence.** Our results show that performance can shift notably depending on the evaluation protocol (e.g., direct selection vs. entailment scoring). Future work should explore hybrid or adaptive evaluation strategies that better capture pragmatic reasoning under uncertainty.

Ethics Statement

This research utilizes publicly available models and visual assets. Human evaluations were conducted with compensated undergraduate students who provided informed consent and were trained on the task. No personal or sensitive data were collected, and no human participants were directly involved beyond voluntary evaluation of anonymized examples. We acknowledge that the models may reflect societal biases from their training data, which could manifest in clue generation or interpretation. Therefore, we caution against deploying automated abductive reasoning systems in sensitive contexts—such as psychological profiling or cultural analysis—without robust fairness auditing and mandatory human oversight to mitigate potential harms.

Acknowledgments

The authors of this paper were supported by the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China.

References

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. 2025. Llm-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. Preprint, arXiv:2310.03903.

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.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. Preprint, arXiv:1908.05739.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. Preprint, arXiv:1911.11641.

Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the utility of learning about humans for human-ai coordination. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Harry Frankfurt. 1958. Peirce's notion of abduction. Journal of Philosophy, 55:593–596.

Google. 2025. Gemini: A family of highly capable multimodal models. Preprint, arXiv:2312.11805.

Jiaxian Guo, Bo Yang, Paul Yoo, Bill Yuchen Lin, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Suspicion-agent: Playing imperfect information games with theory of mind aware gpt-4. Preprint, arXiv:2309.17277.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. Preprint, arXiv:2202.04800.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 9118–9147. PMLR.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. 2022. Visual abductive reasoning. Preprint, arXiv:2203.14040.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. Preprint, arXiv:2304.08485.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. Preprint, arXiv:2004.10796.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST '23, New York, NY, USA. Association for Computing Machinery.

Charles Sanders Peirce. 1931. Collected Papers of Charles Sanders Peirce, Vol. 5: Pragmatism and Pragmaticism. Harvard University Press.

Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W. Malone, and Anita Williams Woolley. 2021. Quantifying collective intelligence in human groups. Proceedings of the National Academy of Sciences, 118.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. Preprint, arXiv:2503.19786.

Rose E. Wang, Sarah A. Wu, James A. Evans, David C. Parkes, Joshua B. Tenenbaum, and Max Kleiman-Weiner. 2021. 152too many cooks: Bayesian inference for coordinating multi-agent collaboration. In Human-Like Machine Intelligence. Oxford University Press.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. Preprint, arXiv:2206.07682.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. Preprint, arXiv:2305.10601.

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. Preprint, arXiv:2306.05685.

Tianshi Zheng, Jiayang Cheng, Chunyang Li, Haochen Shi, Zihao Wang, Jiaxin Bai, Yangqiu Song, Ginny Y. Wong, and Simon See. 2025. Logidynamics: Unraveling the dynamics of logical inference in large language model reasoning. Preprint, arXiv:2502.11176.

A Related Work

LLMs in Multi-Agent Environments. Recent research has increasingly explored the use of Large Language Models (LLMs) as autonomous agents capable of reasoning and planning in interactive environments (Zheng et al., 2023; OpenAI et al., 2024; Wei et al., 2022; Yao et al., 2023). When augmented with capabilities like memory, belief modeling, or tool usage, LLMs can perform complex multi-step reasoning, often outperforming traditional reinforcement learning methods in settings such as open-domain survival games and two-player imperfect information games (Park et al., 2023; Huang et al., 2022; Guo et al., 2024). Another line of work has focused on multi-agent coordination, using both simulated environments (Carroll et al., 2019; Wang et al., 2021; Agashe et al., 2025; Li et al., 2023) and models of human collaboration (Riedl et al., 2021). While this body of research demonstrates LLMs' potential for strategic planning, it has primarily centered on tasks with clear, instrumental objectives. **In contrast, our work investigates the more nuanced challenge of creative communication, where agents must convey and interpret ambiguous concepts under imperfect information.**

Abductive Reasoning. Abductive reasoning, the process of generating and verifying hypotheses to best explain observations (Peirce, 1931), is a long-standing challenge in AI. In NLP, this ability has been predominantly studied through static, single-turn benchmarks. Early works such as aNLI (Bhagavatula et al., 2020) and PIQA (Bisk et al., 2019) focused on text-based abductive inference, which was later extended to the visual domain with datasets like VisualCOMET (Park et al., 2020). More recent approaches have explored structured abductive inference to improve logical consistency in LLMs (Zheng et al., 2025). **However, these single-agent paradigms fail to capture the dynamic, communicative nature of abduction as it occurs in collaborative problem-solving.** Our work addresses this gap by introducing an interactive, multi-agent setting that jointly evaluates both hypothesis generation and selection under adversarial conditions.

B Model Details

In our main experiments, we evaluated 6 VLMs. An additional model (Qwen2.5-VL-72B) is included for scaling analysis in Appendix I.

- **Qwen2.5-VL-7B / Qwen2.5-VL-32B / Qwen2.5-VL-72B** (Bai et al., 2025) are open-source multimodal large language models developed by Alibaba, supporting both text and vision understanding with strong scaling performance across different parameter sizes. The 72B variant is included to verify whether the storyteller–listener asymmetry persists at frontier open-source scale.
- **Gemma3-12B / Gemma3-27B** (Team et al., 2025) are the latest generation of Google's Gemma open-source models, designed with improved efficiency and performance for both language and multimodal reasoning tasks.
- **Gemini-2.5-Flash** (Google, 2025) is a proprietary multimodal model from Google, optimized for efficiency and fast inference, while maintaining competitive reasoning ability.
- **GPT-4o** (OpenAI et al., 2024) is OpenAI's flagship multimodal model, supporting joint reasoning over text and images with strong listener performance in our evaluation.

All models are accessed through the OpenRouter API, with temperature set to 0.7, with total cost of approximately $100 USD.

C Prompt Details

Storyteller: Select Target Image

```
You are a storyteller in a Dixit game. You must select
one card from your 4-card hand as the target image. Your
goal: maximize your own score by ensuring some, but not
all, players guess correctly.
IMPORTANT: Respond strictly in JSON format:

    {
        "reasoning": "Brief analysis (max 50 words)",
        "answer": "The card number (1-4)"
    }
```

Storyteller: Generate Description

```
You are a storyteller in a Dixit game. Create a
description for your chosen image. Scoring rules: -
All guess correctly = 0 points - None guess correctly
= 0 points - Some guess correctly = 3 points (optimal)
Your description should balance ambiguity and clarity,
using metaphorical and emotional language.
IMPORTANT: Respond strictly in JSON format:

    {
      "reasoning": "Reasoning towards final description",
        "answer": "Your crafted Dixit description"
    }
```

Listener: Select Distractor Image

```
You are a player in a Dixit game. Given the storyteller's description, choose one card from your hand that could plausibly match it (but is not the target). Your goal: mislead others into choosing your card.
IMPORTANT: Respond strictly in JSON format:
    {
        "reasoning": "Brief analysis (max 50 words)",
        "answer": "The card number (1-4)"
    }
```

Listener: Direct Selection Strategy

```
You are a player in a Dixit game trying to guess which image matches the storyteller's description. Evaluate all candidates and select the one that best fits.
IMPORTANT: Respond strictly in JSON format:
    {
        "reasoning": "Brief analysis (max 50 words)",
        "answer": "The candidate number (1-N)"
    }
```

Listener: Entailment Scoring Strategy

```
You are evaluating how well an image matches a given clue in a Dixit game. Provide a numerical score (0-100) with reasoning.
IMPORTANT: Respond strictly in JSON format:
    {
       "reasoning": "Detailed reasoning for the score",
        "answer": "Your numerical rating (0-100)"
    }
```

D DixitArena Environment Details

Data and Card Pool. We use a custom-made set of 84 Dixit-style illustrations (`1.png`–`84.png`), stored under `images/`. Each image is in PNG format (305×460, RGBA, 8-bit per channel) and transmitted to the API via Base64 encoding without preprocessing. In each match, four players are dealt 4 cards each from a shuffled pool; in every round, the storyteller selects one target and the three guessers each contribute one distractor. These four images are shuffled using `random.shuffle()` to form the candidate set. Guessers select among them (excluding their own card), with distractor sampling guided by each model's semantic matching rather than random choice. This ensures diversity and realism in distractor quality.

Game Flow. The game proceeds in rounds: (i) the storyteller selects a target and produces a clue, (ii) the other players submit distractors, (iii) the candidate set is shuffled and displayed, (iv) guessers select a card, and (v) scoring is applied. The scoring rules are: partial correct → storyteller +3 and correct guessers +3; all correct → storyteller 0, guessers +3; all wrong → storyteller 0, guessers +2; any distractor chosen → card owner +1. Each match lasts 24 rounds (two phases of 12). In Phase 2, players swap hands (P1↔P3, P2↔P4) to eliminate hand-quality bias. The evaluation spans 21 matches in a full round-robin, including self-play.

Randomness and Reproducibility. We fix the random seed at 42 for shuffling and candidate order. Single-run evaluation produces 504 rounds in total (21 matches × 24 rounds). All rounds, decisions, and scores are recorded in structured JSON, ensuring exact reproducibility given the same seed and configuration.

Fairness Guarantees. Fairness is achieved through hand-swap across phases, equal storyteller turns for all players, candidate shuffling to prevent position bias, and full round-robin coverage (including self-play) to eliminate matchup imbalance.

Open Source Resources. We release code, configuration files, and sample logs for reproducibility at `https://github.com/HKUST-KnowComp/DixitWorld`. A minimal demo script (`simple_demo.py`) reproduces a 2-round match for quick verification.

E DixitBench Curation Details

DixitBench serves as an auxiliary benchmark that evaluates how well models can discriminate semantically similar yet subtly distinct visual descriptions under controlled difficulty levels.

Data Curation Pipeline. We first curate captions for all 84 Dixit-style illustrations using a vision-language model prompted to *"describe this artwork implicitly with a single abstract phrase."* Each caption represents a mid-level abstraction between a single-word concept and a full-sentence description, reflecting the ambiguous, poetic quality typical of Dixit gameplay.

Embedding and Similarity Computation. All captions are encoded into 384-dimensional semantic embeddings using the `sentence-transformers/all-MiniLM-L6-v2` model. Pairwise cosine similarities are computed to produce an 84×84 similarity matrix. This matrix serves as the foundation for controlled distractor sampling based on semantic closeness.

Difficulty-Based Distractor Sampling. For each target image i, we identify distractor candidates by ranking all other images according to cosine similarity:

- **Hard:** top 5 most similar captions (ranks 1–5)
- **Medium:** randomly sampled from ranks 10–20
- **Easy:** randomly sampled from ranks 30–80

This design creates three difficulty tiers spanning high- to low-confusability distractor sets. Each target yields three distractor groups, resulting in **252 benchmark items** (84 images $\times$ 3 difficulty tiers), each with 1 target and 5 distractors (6 candidates total).

Human Evaluation. To validate benchmark quality, we conducted human annotation on 30 sampled items (10 from each of the Easy, Medium, and Hard tiers). Annotators rated (i) caption-image coherence, (ii) distractor plausibility, and (iii) difficulty consistency. Average agreement exceeded 0.8 (Cohen's κ), confirming the benchmark's semantic validity.

This dataset complements DixitArena by isolating the perception component of abductive reasoning—testing whether a model can distinguish subtle conceptual differences in metaphorical phrasing before interacting in a multi-agent setting. The phrase-level dataset and scripts will be released alongside our main benchmark to facilitate reproducibility.

F DixitBench Similarity Statistics

We report the distributional properties of the caption-based similarity scores used to construct DixitBench distractor sets.

Across all $\binom{84}{2} = 3{,}486$ image pairs, pairwise cosine similarities range from -0.147 to 0.655 (mean $= 0.218$, median $= 0.215$, $\sigma = 0.108$). Distractor sampling yields well-separated difficulty tiers:

Difficulty	Mean Sim.	Min	Max
Easy	0.171	-0.049	0.322
Medium	0.307	0.117	0.432
Hard	0.414	0.208	0.655

Table 4: Average distractor–target similarity by difficulty tier. Hard distractors are 2.4$\times$ more similar to targets than Easy ones.

The clear separation between tiers confirms that our similarity-based sampling produces meaningfully different difficulty levels. The full 84×84 similarity matrix, embedding vectors, and sampling code are included in the supplementary release.

G Computation of the Clarity Metric

For transparency, we provide the exact computation procedure for the "Clarity" metric used in human evaluation. This metric measures how well a storyteller's clue balances ambiguity and precision, penalizing both overly vague and overly literal descriptions. The computation follows a three-step normalization and penalty process:

1. **Linear transformation:** Each raw rating s (from 1 to 5) is mapped to a centered clarity score $S^\star$:

$$S^\star = 1 - \frac{|s-3|}{2}, \quad s \in \{1,2,3,4,5\}. \tag{2}$$

   This yields $S^\star = 1.0$ for $s = 3$ (perfect balance), 0.5 for $s = 2, 4$, and 0.0 for $s = 1, 5$.

2. **Rater aggregation:** For each clue, multiple raters' scores are aggregated by taking the median of their $S^\star$ values:

$$\text{Clarity}_{clue} = \text{median}(S_1^\star, \ldots, S_n^\star). \tag{3}$$

   Using the median rather than the mean prevents misleading "false middle" effects (e.g., half of raters giving 1 and half giving 5).

3. **Extreme penalty:** To penalize polarized judgments, we down-weight the clarity score by the fraction of extreme votes:

$$S^{final} = \text{Clarity}_{clue} \times \Big(1 - \frac{\#\{s=1 \vee s=5\}}{n}\Big). \tag{4}$$

   This ensures that clues judged simultaneously "too vague" and "too obvious" receive low final scores even if their median appears moderate.

The resulting S^{final} value is averaged across all evaluated clues to yield the "Clarity" score reported in Table 3.

H Human Evaluation Protocol

We provide details of the human evaluation procedure for transparency and reproducibility.

Annotators. Five compensated undergraduate students (all fluent in English) were recruited and trained on the evaluation task. Each annotator completed a 15-minute training session with example Dixit clues and scoring rubrics before independent annotation.

Task Design. For each evaluation item, annotators were shown: (i) the storyteller's clue text, (ii) the target image, and (iii) the set of candidate images (target + distractors). They rated each clue on two 5-point Likert scales:

- **Clarity** (1–5): How well does the clue guide a player toward the target? (1 = completely unclear, 5 = immediately obvious)
- **Creativity** (1–5): How imaginative and non-literal is the clue? (1 = purely literal, 5 = highly creative)

Additionally, annotators attempted to identify the target image from the candidate set, providing a direct measure of listener accuracy.

Quality Control. Inter-rater agreement was measured using Cohen's κ. For the benchmark curation task, $\kappa > 0.8$ (near-perfect agreement). For the main clarity/creativity evaluation, $\kappa \approx 0.75$ (substantial agreement). Annotators whose agreement fell below $\kappa = 0.6$ on calibration items were retrained before continuing. The evaluation interface was implemented as an HTML form with randomized item ordering to prevent position bias.

I Scaling Analysis

To isolate the effect of model scale on abductive reasoning, we examine the Qwen2.5-VL family across three sizes (7B, 32B, 72B), which share the same architecture and training methodology.

Model	Story.	Listen.	Overall	Bench
Qwen2.5-VL-7B	14.29	30.87	17.39	36.67
Qwen2.5-VL-32B	15.48	50.79	30.63	63.10
Qwen2.5-VL-72B	4.76	49.21	22.59	61.90

Table 5: Scaling analysis within the Qwen2.5-VL family. "Story."/"Listen." = DixitArena normalized scores (%); "Bench" = DixitBench total accuracy (%). Listener ability scales with size, while storyteller ability does not.

Two clear patterns emerge. First, **listener performance scales with model size**: DixitBench accuracy improves from 36.67% (7B) to 63.10% (32B), with 72B performing comparably at 61.90%. The same trend holds for Arena listener scores (30.87% $\rightarrow$ 50.79% $\rightarrow$ 49.21%). Second, **storyteller performance does not scale**—and in fact decreases at 72B (4.76%), the lowest among all models tested. This striking divergence confirms that the storyteller–listener asymmetry is a structural limitation of current VLM architectures, not a matter of insufficient scale. Scaling improves discriminative hypothesis selection but fails to develop the pragmatic control needed for creative, ambiguity-balanced hypothesis generation.

J Storyteller Failure Mode Diagnostics

To systematically characterize why storytellers fail, we analyzed clue texts from representative DixitArena matches. Failures fall into two distinct categories:

Over-Specification (All-Correct). When storytellers produce overly literal descriptions, all listeners identify the target trivially, yielding zero points. Example: GPT-4o generated *"Tiny warriors battling for a golden throne"* for an image of small figurines near a crown—every listener matched it immediately. These clues lack the abstraction needed for productive ambiguity.

Semantic Mismatch (All-Wrong). Conversely, when clues fail to connect with the target's visual semantics, no listener succeeds. Example: Gemma3-12B produced *"A slow climb to nowhere"* for an image that depicted an unrelated scene—the metaphor was disconnected from any visual element, leaving all listeners guessing incorrectly.

Successful Ambiguity (Partial-Correct). Effective clues balance metaphor with grounding. Example: Gemma3-27B's *"The last attempt to hold onto a fading dream"* guided 2 of 3 listeners to the correct image while the third was misled by a plausible distractor. The clue's emotional abstraction created productive interpretive space.

These patterns confirm that the storyteller challenge is fundamentally one of pragmatic calibration: clues must be abstract enough to create ambiguity but grounded enough to remain solvable—a narrow corridor that current VLMs struggle to navigate regardless of scale.

K Qualitative Case Studies

This appendix presents representative examples of actual game rounds from our DixitArena evaluation, illustrating both successful and failed storytelling strategies across different models.

K.1 Case 1: Successful Storytelling (Partial-Correct)

Model: Gemma3-27B (Storyteller)
Round: Match 3, Phase 1, Round 3
Clue: "A delicate hope, reaching for something unseen, supported by a fragile network."
Target: Image 35.png
Outcome: 2/3 listeners guessed correctly (Partial-Correct).
Insight: Balanced ambiguity with metaphorical language guided some listeners while misleading others.

K.2 Case 2: Failed Storytelling (All-Correct, Too Obvious)

Model: Qwen2.5-32B (Storyteller)
Round: Match 7, Phase 1, Round 5
Clue: "A knight in shining armor on horseback, holding a spear, as a tentacle rises from the pages of a book."
Target: Image 10.png
Outcome: 3/3 listeners guessed correctly (All-Correct).
Insight: Over-specific literal description eliminated ambiguity, leaving no room for alternative interpretations.

K.3 Case 3: Successful Storytelling (Poetic Ambiguity)

Model: Gemma3-27B (Storyteller)
Round: Match 7, Phase 1, Round 3
Clue: "A fleeting glimpse of a world unseen."
Target: Image 11.png
Outcome: 1/3 listeners guessed correctly (Partial-Correct).
Insight: Poetic abstraction created interpretive space; one aligned listener succeeded while others were misled.

K.4 Case 4: Failed Storytelling (All-Wrong, Too Vague)

Model: Gemma3-12B (Storyteller)
Round: Match 6, Phase 1, Round 3
Clue: "A forgotten trophy, watching over a curious secret."
Target: Image 56.png
Outcome: 0/3 listeners guessed correctly (All-Wrong).
Insight: Semantic mismatch: the clue did not align with the target image, and even the same-model listener failed.

K.5 Overall Insights

Across these cases, we observe distinct success and failure patterns. Successful clues (Cases 1 and 3) typically rely on metaphorical or poetic language that balances clarity and ambiguity, allowing some listeners to be guided while others are misled. In contrast, failures arise from two opposite directions: over-specification (Case 2), where excessive literal detail leaves no room for interpretation, and semantic mismatch (Case 4), where vague or misaligned clues fail to connect with the intended target. These qualitative examples complement the aggregate statistics by revealing *why* models succeed or fail in abductive storytelling.

L Per-Image Category Analysis

To investigate whether model performance varies systematically across image types, we classified the 84 Dixit illustrations into *metaphorical* (53 images) and *literal* (31 images) categories based on their visual content. Results include Qwen2.5-VL-72B from the scaling study.

Model	Metaphorical	Literal	Gap
Qwen2.5-VL-32B	67.3%	66.7%	+0.6
Qwen2.5-VL-72B	67.3%	61.3%	+6.0
Gemma3-12B	64.2%	64.5%	−0.4
Gemma3-27B	56.0%	62.4%	−6.4
Gemini-2.5-Flash	69.2%	61.3%	+7.9
GPT-4o	**80.5%**	63.4%	**+17.1**

Table 6: DixitBench accuracy by image category. GPT-4o shows a pronounced advantage on metaphorical scenes (+17.1pp), while Gemma3-27B performs slightly better on literal scenes.

GPT-4o's strong advantage on metaphorical images (+17.1pp gap) suggests superior ability to process abstract, non-literal visual semantics. Conversely, Gemma3-27B shows a reversed pattern (−6.4pp), performing better on literal scenes. These asymmetries align with the broader finding that proprietary models excel at nuanced hypothesis selection, while open-source models may rely more on concrete visual features.

M Additional Results

M.1 Evaluation under Entailment Scoring

To further assess the robustness of listener evaluation, we additionally tested models on DixitBench using an alternative entailment scoring scheme (Table 7). Unlike the standard direct selection setting where models must choose a single best-matching

image, the entailment variant requires assigning independent plausibility scores (0–100) to all candidates. This approach evaluates a model's ability to produce well-calibrated absolute judgments rather than comparative preferences.

Overall, results show that entailment scoring yields moderately lower accuracies across all models—particularly on the Hard subset—while preserving the same relative ranking trends observed under direct selection. GPT-4o and Gemini-2.5-Flash remain the strongest performers, suggesting that although VLMs handle relative choice well, their absolute entailment calibration remains underdeveloped. This supports our earlier finding that current VLMs are optimized for *comparative abductive reasoning* rather than calibrated entailment verification.

Model	Easy	Hard	Total
Qwen2.5-VL-7B	22.47	18.19	20.33
Qwen2.5-VL-32B	58.37	49.81	54.09
Gemma3-12B	56.03	40.51	48.27
Gemma3-27B	58.23	50.41	54.32
Gemini-2.5-Flash	61.21	53.43	57.32
GPT-4o	**66.69**	**52.41**	**59.55**

Table 7: Performance on *DixitBench* under the entailment scoring scheme. Scores are moderately lower than under direct selection, especially on the Hard subset, while relative ranking remains consistent.

M.2 Leave-One-Listener-Out Results

Table 8 reports the leave-one-listener-out (LOLO) analysis. For each model, we compare the original storyteller score with the average score when one listener is removed at a time. The small average differences and high stability indices confirm that no single listener dominates the outcome.

Model	Orig. Score	Avg Δ	Std Δ	Stability
Qwen2.5-VL-7B	21	-0.08	0.28	0.908
Qwen2.5-VL-32B	27	-0.11	0.31	0.897
Gemma3-12B	60	-0.24	0.43	0.858
Gemma3-27B	72	-0.29	0.45	0.849
Gemini-2.5-Flash	69	-0.27	0.45	0.851
GPT-4o	51	-0.20	0.40	0.866

Table 8: Leave-one-listener-out (LOLO) storyteller scores. Stability is defined as $1 - \text{Std}(\Delta)/3.0$, ranging from 0–1 (higher = more robust).

N DixitBench Proxy Robustness

To verify that DixitBench's strong correlation with DixitArena listener performance ($r = 0.947$) is not driven by any single model, we perform a leave-one-out cross-validation (LOOCV) analysis.

Model Dropped	Pearson r
None (full, n=7)	0.947
Qwen2.5-VL-7B	0.770
Qwen2.5-VL-32B	0.947
Qwen2.5-VL-72B	0.947
Gemma3-12B	0.960
Gemma3-27B	0.947
Gemini-2.5-Flash	0.956
GPT-4o	0.986

Table 9: Leave-one-out Pearson correlation between DixitBench and DixitArena listener scores. The correlation remains strong ($r \geq 0.770$) across all subsets.

Dropping any single model yields $r \geq 0.770$, with 6 of 7 subsets maintaining $r \geq 0.947$. The only notable decrease occurs when removing Qwen2.5-VL-7B, which acts as a high-leverage point due to its uniformly low performance across both benchmarks. This LOOCV analysis confirms that DixitBench is a robust proxy for Arena listener performance, not an artifact of any particular model's behavior.

O Position Selection Bias

To verify that model performance is not confounded by positional preferences, we analyze predicted position distributions across all 252 DixitBench items (all three difficulty tiers, with 6 candidate positions per item; including Qwen2.5-VL-72B from the scaling study). While Section 3.4 reports aggregate uniformity ($\chi^2 = 1.83, p = 0.61$), per-model analysis reveals notable variation.

Model	χ^2	P1	P2	P3	P4	P5	P6
Qwen2.5-VL-32B	5.0	33	38	38	49	47	47
Gemini-2.5-Flash	15.2	56	44	26	37	36	53
Qwen2.5-VL-72B	18.5	35	25	38	42	52	60
Gemma3-27B	35.2	67	28	27	24	32	39
GPT-4o	57.7	36	30	33	45	46	61
Gemma3-12B	67.6	72	36	35	30	39	37

Table 10: Position prediction distribution (count out of 252 samples) and χ^2 goodness-of-fit statistic. Lower χ^2 indicates more uniform selection. Qwen2.5-VL-32B shows the most uniform behavior; Gemma3-12B shows the strongest bias toward Position 1.

Two patterns emerge: (i) smaller models exhibit stronger positional biases (Gemma3-12B: $\chi^2 = 67.6$, favoring Position 1), while mid-to-large models are more uniform; and (ii) accuracy varies by correct-answer position (Position 5: 76.3%, Position 2: 59.3%), suggesting that candidate ordering interacts with model attention patterns. Crucially, our hand-swap and shuffling procedures ensure that

these per-model biases do not systematically advantage any model in aggregate comparisons.

P Calibration Analysis

To assess confidence calibration under entailment scoring, we compute Expected Calibration Error (ECE) and Brier scores for each model. This analysis uses the subset of entailment-scoring runs for which per-candidate score vectors were logged (smaller than the full entailment evaluation reported in Table 7); accuracy numbers here therefore differ slightly from that table. The entailment scoring strategy assigns plausibility scores (0–100) to each candidate; we convert these to probabilities via softmax normalization and measure the gap between confidence and accuracy.

Model	Diff.	Acc.	ECE ↓	Brier ↓
GPT-4o	Easy	60.7%	0.204	0.223
GPT-4o	Hard	53.6%	0.242	0.251
Gemini-2.5-Flash	Easy	57.1%	0.251	0.275
Gemini-2.5-Flash	Hard	52.4%	0.270	0.280
Gemma3-27B	Easy	53.6%	0.276	0.293
Gemma3-27B	Hard	41.7%	0.186	0.268
Gemma3-12B	Easy	60.7%	0.309	0.318
Gemma3-12B	Hard	41.7%	0.171	0.254

Table 11: Calibration metrics under the entailment scoring scheme. Lower ECE and Brier scores indicate better-calibrated confidence estimates.

GPT-4o exhibits the best overall calibration (lowest Brier scores across both difficulty levels), consistent with its superior entailment performance. However, all models show substantial ECE values (> 0.17), indicating a persistent gap between expressed confidence and actual accuracy. This confirms that current VLMs, while reasonably accurate at comparative selection, remain poorly calibrated for absolute plausibility judgments.

Q Sensitivity Analysis

To verify the robustness of our findings, we conducted a seed sweep using three random seeds (42, 100, 200) for two representative models on DixitBench.

Model	Seed	Easy	Hard	Total
GPT-4o	42	81.0	72.6	76.8
GPT-4o	100	75.0	69.0	72.0
GPT-4o	200	79.8	73.8	76.8
Gemma3-27B	42	75.0	67.9	71.4
Gemma3-27B	100	79.8	72.6	76.2
Gemma3-27B	200	76.2	65.5	70.8

Table 12: DixitBench accuracy (%) across three random seeds. Standard deviations remain within $\pm$3.0pp.

Both models show low variance ($\leq$3.0pp), confirming that our evaluation is robust to random seed choice. The absolute accuracy levels here differ from Table 2 because this is an independent re-run with fresh candidate shuffling; what matters for robustness is the low within-seed variance shown above. Additionally, a prompt-style ablation (poetic vs. concrete phrasing) shifts metrics by less than 5 percentage points, confirming that our findings are not sensitive to specific prompt formulations.

R Entailment Strategy Error Analysis

To understand why entailment scoring underperforms direct selection (Section 3.2), we analyze the score distributions produced by each model. In the entailment scheme, models assign plausibility scores (0–100) to each candidate independently.

Model	Acc.	Marg.	OC%
GPT-4o	57.1	+5.9	42.3
Gem.-Flash	54.8	+2.6	45.2
Gem3-12B	51.2	+0.2	48.8
Gem3-27B	47.6	−0.1	52.4

Table 13: Entailment error analysis. Marg. = correct − best wrong score; OC = overconfident wrong predictions ($\geq$70).

Across all models, 42–52% of errors involve high-confidence wrong predictions (score $\geq$70 for an incorrect candidate). Gemma3-27B shows the worst calibration: its average margin is effectively zero (-0.1), meaning the correct answer is no more likely to receive the highest score than a distractor. On the Hard subset, margins become negative for both Gemma models (-1.4 and -1.6), indicating that increased distractor similarity causes systematic overscoring of wrong candidates. This overconfidence on distractors is the primary failure mode of entailment scoring, explaining its consistent underperformance relative to comparative direct selection.

S Temperature Sensitivity

We evaluate the effect of sampling temperature on DixitBench listener accuracy. This is an independent re-run (on the Easy+Hard subset, 168 items) with fresh random candidate shuffling, so absolute numbers may differ from Table 2; we focus on the relative trend across temperatures.

Model	t=0.3	t=0.5	t=0.7	t=1.0
GPT-4o	69.6%	64.9%	**73.8%**	60.7%
Gemma3-27B	74.4%	**76.8%**	57.7%	73.2%

Table 14: DixitBench accuracy across sampling temperatures. Optimal temperature differs by model: t=0.7 for GPT-4o, t=0.5 for Gemma3-27B.

Temperature sensitivity varies substantially across models. GPT-4o performs best at t=0.7 and degrades sharply at t=1.0 (-13.1pp), consistent with increased output randomness impairing discriminative reasoning. Gemma3-27B shows a different pattern: it peaks at t=0.5 (76.8%) and drops at t=0.7 (-19.1pp), suggesting that its default configuration may not be optimal for this task. The finding that each model has a distinct optimal temperature highlights the importance of per-model tuning in benchmark evaluation, and supports our decision to use a fixed temperature (t=0.7) for fair cross-model comparison rather than per-model optimization.

T Inference Cost Analysis

For practical deployment, we report the accuracy–cost trade-off measured in average output tokens per DixitBench sample. Accuracies here come from the cost-analysis run (aggregated over all 252 Easy+Medium+Hard items) and may differ slightly from the 168-item Easy+Hard average reported in Table 2.

Model	Acc. (%)	Avg Tokens	Efficiency
Gemini-2.5-Flash	66.3	52	**127.7**
Gemma3-12B	64.3	63	101.9
GPT-4o	74.2	77	97.1
Qwen2.5-VL-72B	65.1	85	77.0
Qwen2.5-VL-32B	67.1	91	73.9
Gemma3-27B	58.3	85	68.8

Table 15: Accuracy vs. inference cost. Efficiency = Accuracy / Avg Tokens $\times$ 100. Gemini-2.5-Flash achieves the best cost–accuracy trade-off; GPT-4o leads in absolute accuracy.

Gemini-2.5-Flash and GPT-4o form the Pareto frontier: Gemini-2.5-Flash offers the best efficiency (127.7) with competitive accuracy, while GPT-4o achieves the highest accuracy at moderate cost. Notably, the Qwen models use substantially more tokens per sample despite comparable or lower accuracy, suggesting less concise reasoning chains. This analysis provides practical guidance for selecting models in resource-constrained evaluation scenarios.