# Saying the Unsaid: Revealing the Hidden Language of Multimodal Systems Through Telephone Games

**Juntu Zhao**
Shanghai Jiao Tong University
Shanghai, China
arossoneri@sjtu.edu.cn

**Jialing Zhang**
Shanghai Jiao Tong University
Shanghai, China
jialingzhang@sjtu.edu.cn

**Chongxuan Li**
Renmin University of China
Beijing, China
chongxuanli@ruc.edu.cn

**Dequan Wang***
Shanghai Jiao Tong University
Shanghai, China
dequanwang@sjtu.edu.cn

## Abstract

Recent closed-source multimodal systems have made great advances, but their hidden language for understanding the world remains opaque because of their black-box architectures. In this paper, we use the systems' preference bias to study their hidden language: During the process of compressing the input images (typically containing multiple concepts) into texts and then reconstructing them into images, the systems' inherent preference bias introduces specific shifts in the outputs, disrupting the original input concept co-occurrence. We employ the multi-round "telephone game" to strategically leverage this bias. By observing the co-occurrence frequencies of concepts in telephone games, we quantitatively investigate the concept connection strength in the understanding of multimodal systems, i.e., "hidden language." We also contribute Telescope, a dataset of 10,000+ concept pairs, as the database of our telephone game framework. Our telephone game is test-time scalable: By iteratively running telephone games, we can construct a global map of concept connections in multimodal systems' understanding. Here we can identify preference bias inherited from training, assess generalization capability advancement, and discover more stable pathways for fragile concept connections. Furthermore, we use Reasoning-LLMs to uncover unexpected concept relationships that transcend textual and visual similarities, inferring how multimodal systems understand and simulate the world. This study offers a new perspective on the hidden language of multimodal systems and lays the foundation for future research on the interpretability and controllability of multimodal systems.

## 1 Introduction

Recent multimodal systems, particularly closed-source ones [Hurst et al., 2024, StepFun, 2024, Bai et al., 2025], have made significant advances, e.g., the newest GPT-4o with Image Generation [OpenAI, 2025] (abbreviated as GPT-4o-IG(20250325)). However, because of these systems' closed features, closed data, and even closed architectures, we are unable to study the systems' understanding of the world using methods based on training. Therefore, test-time methods are urgently needed.

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

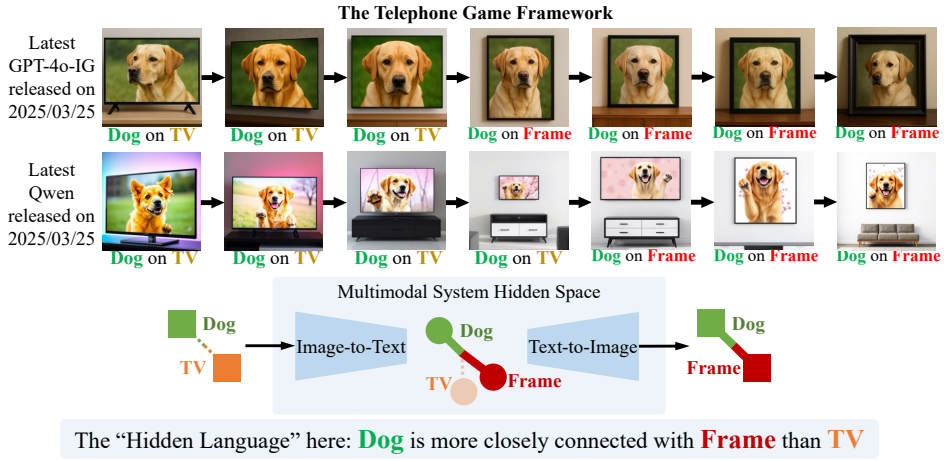



Figure 1: Example 5-round telephone games using the latest SOTA multimodal systems released on 2025.3.25. In each image reconstruction, the system prefers stronger nearby concept connections in multimodal systems' understanding, then changing the outputs. (Extended results on this example are provided in Appendix B), and more examples can be found in Appendix D

The hidden language reflects the connection strength between concepts within multimodal systems [Chefer et al., 2023], offering insight into how they understand the world. While prior training-based methods explored it via internal features [Chen et al., 2023, Chefer et al., 2023, Ghandeharioun et al., 2024], the rise of closed-source models renders such access impossible. Hence, we investigate the hidden language of multimodal systems at test time.

We innovatively propose to strategically leverage the multimodal systems' preference bias to study their hidden language at test time. Multimodal systems are trained to fit textual and visual representations of the same scenes, which typically involves multiple interrelated concepts. Sufficient training strengthens these concept connections in systems' hidden understanding space (abbreviated as hidden space), while limited training weakens them. Therefore, imbalanced training data brings different concept connection strengths, i.e., hidden language. As illustrated in Figure 1, during image-to-text compression, the systems prefer to discard weakly connected concepts; during text-to-image reconstruction, the systems prefer strongly connected concepts [Zhao et al., 2024], even with the latest SOTA GPT-4o-IG(20250325). These preference biases will lead to changes in input concepts, thereby disrupting their co-occurrence in the output scene.

In this paper, we innovatively propose a **test-time** framework based on multi-round **telephone game** to leverage this preference bias, a plug-and-play method involving multiple cycles of image reconstruction. As the telephone game progresses, fragile concept combinations gradually degrade, revealing their fragile connection strength in systems' understanding. And we quantify the connection strength (i.e., hidden language) using the concept co-occurrence frequency in the telephone game. As shown in Figure 2, a higher co-occurrence frequency indicates a stronger concept connection. This metric captures both the training bias and generalization capability: Stronger generalization enables consistent responses to similar patterns, corresponding to a uniform connection strength distribution.

We also contribute Telescope, a dataset consisting of 10,000+ concept pairs derived from 150 common visual concepts, primarily covering basic spatial relations (e.g., "$A$ adjacent to $B$") and some complex interactions (e.g., "$A$ displayed on TV screen"). Leveraging the telephone game and Telescope, we propose a scalable test-time probing framework for the hidden language of multimodal systems: Each new telephone game iteration tends to reveal new concept connections, and as test-time compute scales up, we progressively build a detailed "world map" of the multimodal hidden language.

In this way: **(1)** We uncover key terms associated with a concept in multimodal systems' understanding, revealing the training bias (which combinations are better-trained or not) and the systems' generalization capability; **(2)** By analyzing connection strengths across multiple pathways, we can identify intermediate concepts to enhance concept connections to promote the co-occurrence of discordant concepts; **(3)** Reasoning-LLMs help to understand how the these connection strengths interprets physical-world laws, revealing unexpected relationships beyond textual and visual similarities.

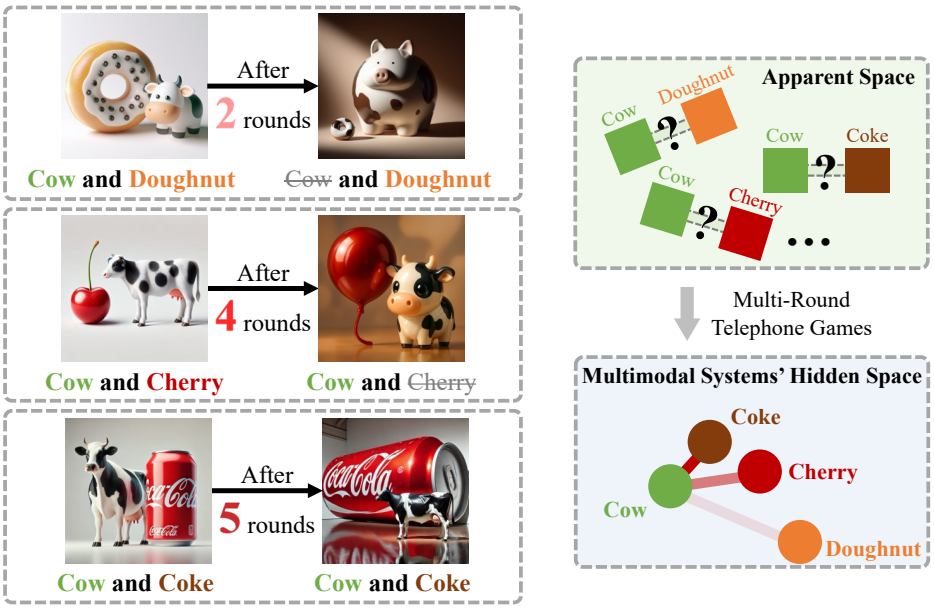



Figure 2: The longevity of concepts combinations in the telephone game (i.e., their co-occurrence frequency) quantitatively reflects the concept connections in multimodal systems' hidden space, termed the "hidden language." (Lighter color means the weaker connection)

**Here, we summarize our contributions:**

- **Test-time Telephone Game Framework**: We innovatively propose to reveal the hidden language of multimodal systems using the framework of telephone game and the concept co-occurrence frequency at test time;
- **Telescope Dataset**: We contribute the Telescope, a database for systematic telephone game probing on multimodal systems' hidden language;
- **Test-Time Scalable Framework**: We keep on creating an increasingly comprehensive hidden language world map of multimodal systems in a scalable way.

## 2 Related Works

**MultiModal Systems** Recent advances in multimodal intelligence systems [Lu et al., 2019, Baltrušaitis et al., 2018, Xie et al., 2024, Guo et al., 2019, Li et al., 2023, 2022, Tan and Bansal, 2019] have shown great ability in processing cross-modal information. Modular pipeline frameworks and autoregressive systems represent two typical paradigms in multimodal architecture: the former leverages V-LLMs [Hurst et al., 2024, Alayrac et al., 2022, Liu et al., 2023, Wu et al., 2024] as core components to construct complex cross-modal connections, while the latter unifies different modalities through sequential modeling within a shared hidden space [Team, 2024, Chern et al., 2024, Li et al., 2025]. Furthermore, in recent years, the emergence of GPT-4o [Hurst et al., 2024] marks a shift toward a fully black-box system paradigm, particularly with the latest version that integrates image generation ability into GPT-4o [OpenAI, 2025]. However, as multimodal systems' internal structures become more complex and black-box, making the way they understand the world and their preferences harder to interpret.

**Hidden Language** In traditional machine learning, researchers could intuitively examine a model's hidden language using tools like attention maps [Vaswani et al., 2017], or Principal Component Analysis(PCA) [Hotelling, 1933]. As deep learning and large-scale models emerge, it becomes common to train lightweight probing models on embeddings to better understand internal representations [Alain and Bengio, 2016, Chefer et al., 2023, Ghandeharioun et al., 2024, Derby et al., 2018, Chen et al., 2023, Frank et al., 2021]. However, with the emergence of many closed-source systems [OpenAI,

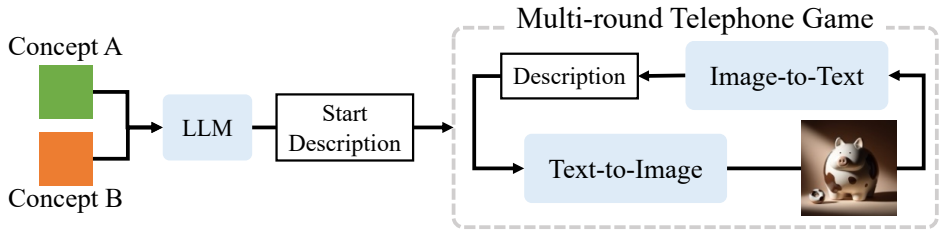



Figure 3: The workflow of telephone game. LLMs convert concept pairs into the start description for telephone game. Then it enters the cycle of text-to-image and image-to-text.

2025] today, we no longer have access even to basic token-level representations. As a result, we are forced to infer the hidden language of multimodal systems directly from their apparent-level outputs, e.g., textual or visual outputs. This constraint effectively pivots the research paradigm from direct internal inspection to a form of behavioral analysis, where the model's observable responses to carefully crafted inputs become the primary source of evidence for its representation structures.

## 3 Framework: Telephone Game

This section introduces our test-time telephone game framework for uncovering the hidden language of multimodal systems, the concept co-occurrence frequency metric for quantifying hidden language, and the Telescope dataset used to support systematic evaluation.

### 3.1 Telephone Game

We propose to use telephone game to study the concept connections in multimodal systems' hidden space, revealing the hidden language. Our telephone game framework involves two key processes:

- Image to Text: When compressing images into text, systems prefer to read more strongly connected concepts in their understanding over those strictly faithful to visual facts. For example in Figure 2, it reads a cow as a pig;
- Text to Image: When reconstructing text to images, systems prefer to create more strongly connected concepts understood from the text to synthesize the visual output. For example, in Figure 2, it creates a balloon instead of a cherry.

These two preferences introduce the different concept connection strengths in multimodal systems, representing the systems' hidden language. To reveal this hidden language at test time, without accessing model parameters, we link the above processes and use changes in apparent space (e.g., image descriptions) to explore the concept connection strengths in the hidden space, as shown in Figure 3. Changes in a single reconstruction may not be apparent in the observable space, e.g., generating a visual resembling both a cow and a pig in Figure 2. Moreover, concepts with fragile connections (rather than absent) may not exhibit a crash initially, but as the cycle progresses, the resulting offsets gradually become apparent. Given these issues, we naturally use the multi-round telephone game to amplify the changes.

In our experiments, for fully integrated multimodal systems like the latest GPT-4o [OpenAI, 2025], we directly utilize the system to perform both text-to-image and image-to-text processes. For multimodal systems composed of separate components, we assemble them using V-LLMs and text-to-image models from the same institution, treating all components as a unified system. All of our instruction prompts are available in Appendix G.

### 3.2 Co-occurrence Frequency

Modern multimodal systems rely on text and visuals, where semantic or visual similarity can seemingly reflect the systems' hidden language. However, as shown in Figure 2, the observed concept connection strengths, e.g., cows and coke, contradict this intuition, highlighting the need for a new metric to capture the hidden language more accurately.


Figure 4: Visualization of several telephone game examples, reflecting which concepts are connected strongly or weakly in the hidden space of multimodal systems, as well as the intermediary concepts that build stronger connections. For more results, see Appendix B.

The tendency of concept pairs to co-occur across multi-round telephone game offers key insight into the hidden language of multimodal systems. Therefore, we propose to use the co-occurrence frequency of different concepts in the multi-round telephone game as a direct measure of connection strength in the system's hidden space, reflecting the hidden language. In a $n$-round telephone game, the co-occurrence frequency of the concept pairs "$A$ and $B$" is defined as:

$$F(A,B) = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n}\mathcal{I}_{i,j}(A,B)}{r \times n} \tag{1}$$

where $r$ means we repeat a telephone game for $r$ times, and $\mathcal{I}_{i,j}(A,B)$ represents whether $A$ and $B$ co-occur in the output of the $j$-th round of the $i$-th telephone game judged by LLMs (the instruction prompt is available in Appendix G). In this study, we choose the image description to analyze the concept co-occurrence frequency.

Note that an implicit limitation lies in the number of rounds. When calculating the metric correlation, we exclude pairs with a co-occurrence frequency of 1.0, as we cannot run an infinite-round telephone game to get a true co-occurrence "probability." And in Section 4.3, we demonstrate some interesting phenomena that emerge as the round number increases.

### 3.3 Dataset: Telescope

We also contribute a dataset, Telescope. We collect 150 common concepts as the basic "vocabulary" of the hidden language. In this study, we focus on combinations of 2 concepts, with plans to explore complex combinations of more concepts in future work.

In the main portion of the dataset, these concepts form 11,175 concept pairs, each involving arranging 2 concepts side by side, called the simple-pattern. The other portion of the dataset represents more complex combinations of concepts, called complex-pattern: We investigate 3 interesting visual fusion strategies: displaying Concept $A$ on a TV screen, creating Concept $A$ in the visual style of Van Gogh, and constructing Concept $A$ using wood as a material. Unlike simple-patterns, they involve interactions between different concepts. The Telescope dataset allows us to explore how the system establishes its hidden language in the hidden understanding space.

### 3.4 LLMs as "MLPs"

Concept connections, viewed as the hidden language, not only reveal training data biases and generalization capabilities, but also open the door to deeper inquiry: ***What further insights into the system's hidden logic about how it understands and simulates the real-world physical laws might emerge?*** In this study, we abstract the text as a special "embedding" bridging hidden features and observed pixels. In conventional deep learning, linear probes (e.g., MLPs) interpret embeddings to reveal internal logic; analogously, we employ Reasoning-LLMs as cognitive probes to parse textual evolution across rounds. These analyses uncover implicit constraints on real-world laws beyond observed-level correlations (e.g., textual or visual similarity), suggesting that multimodal systems attempt to simulate human world laws and causal relationships. The experimental details and results can be found in our Appendix C.

# 4 Experiments

## 4.1 Model and Dataset

**Model** Our primary experiments utilize OpenAI's multimodal system, recognized as SOTA. Here, we use the system composed of GPT-4o [Hurst et al., 2024] and Dall·E-3 [OpenAI, 2023], which is also the configuration used in OpenAI's official products. Preliminary results show that even for simple tasks, i.e., a single concept or two identical concepts, after a 5-round telephone game, the original concepts exhibit crashes (disappearing or transforming, might because of the emergence of irrelevant concepts) at rates of 26.4% and 24.4%, respectively, highlighting the significant bias in multimodal systems, forming a key basis for our framework, as analyzed in Section 3. As for the latest GPT-4o-IG(20250325), we present the experimental results using the web version of the tool in Appendix D.

In Section 4.2, we also analyze the hidden language of various multimodal systems derived from different sources, including: (1) StepFun [StepFun, 2024]: Step1V and Step1X, (2) Qwen [Bai et al., 2025]: Qwen2.5-VL and Wanx2.1. Simple open-source systems are excluded, as closed-source systems now far outperform open-source ones. Open-source systems face limitations in knowledge acquisition and input text length. For an extreme example, if the text is as short as only 3 words, e.g., "$A$ and $B$", the exposure of issues would largely depend on chance. But we can use its open-access features to validate multimodal systems' preference for strong concept connections, see Appendix F.

**Dataset** Our dataset, Telescope, consists of over 10,000 concept pairs. Due to time and cost constraints, we present results on a refined subset, with selection strategies detailed in the following sections. However, we will not stop our experiments, and we are committed to continuously expanding the global map of multimodal systems' hidden language.

## 4.2 Correlation

| | | | |
|---|---|---|---|
| Metric Correlation ↑ | Co-occur vs Semantic<br>0.046 | Co-occur vs Visual<br>-0.178 | Semantic vs Visual<br>0.041 |
| System Correlation ↑ | OpenAI vs StepFun<br>0.506 | OpenAI vs QWen<br>0.475 | StepFun vs QWen<br>0.503 |

Table 1: Pearson Correlation Coefficients among the 3 metrics and different multimodal systems. Using the OpenAI system to calculate co-occurrence frequency (Co-occur) for metric comparison and our Co-occur metric for analyzing hidden language correlations across different systems, we find that semantic and visual similarities fail to capture the hidden language, while hidden languages across different systems show good correlation.

The concepts' co-occurrence frequency reflects their connection strength in hidden space. We are particularly interested in whether this phenomenon can be explained by existing similarity metrics, such as semantic and visual similarity. First, we detail the setup:

- Metric: To explore the correlation of different metrics, we use CLIP model [Radford et al., 2021] for semantic embeddings and ResNet-50 [He et al., 2016] for visual embeddings, and compute concepts similarity between embeddings;
- Models: To explore the correlation of hidden languages in different systems, we implement 3 multimodal systems: OpenAI, StepFun, and QWen;
- Dataset: Given the substantial cost, we rank the simple-pattern concept pairs in Telescope by the average of their semantic and visual similarities, and uniformly sample 400 pairs.
- Telephone Game: For each pair, we repeat a 5-round telephone game for 3 times;
- Focus: Here, we focus on the connections between the original input concepts. The new concepts emerging during the telephone game will be analyzed in Section 4.3.

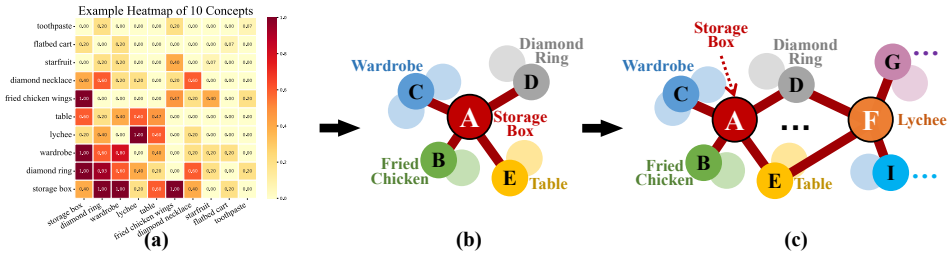



Figure 5: Our scalable framework: (a) Basic Connection: we reveal concept connections and identify nearby keywords; (b) Local Connection: repeated telephone games establish a local graph around a concept; (c) Global Connection: increasing telephone games connect local structures, forming a comprehensive "world map" of the multimodal hidden language!

#### 4.2.1 Metric Correlation

We conduct quantitative correlation analysis to examine relationships between different metrics. Among the 400 concept pairs, 246 experience concept crashes. As discussed in Section 3.2, due to the round number limitation, concepts with a co-occurrence frequency of 1.0 (i.e., no crash) are excluded from metric correlation analysis, as these frequencies might be unreliable. Therefore, the correlation is computed based on the 246 crashed pairs. We also present an intuitive visualization of their correlation in Appendix A

For each of the 246 concept pairs, we compute semantic similarity, visual similarity, and co-occurrence frequency between the 2 concepts in the pair. We calculate the Pearson Correlation Coefficients [Pearson, 1895] by measuring pairwise correlations between the three 246-length lists. As shown in Table 1, semantic and visual similarities fail to capture concept connections, underscoring the need for metrics like our co-occurrence frequency.

#### 4.2.2 System Correlation

We further examine whether the hidden languages of multimodal systems from different sources are correlated. We run the telephone game on 400 concept pairs using StepFun and QWen systems, repeating each experiment 3 times. Since the same metric is compared across systems, the round number limitation does not apply.

As reported in Table 1, we observe a moderate correlation between the hidden languages of different multimodal systems, indicating potentially consistent concept connections in the hidden space. This consistency significantly surpasses correlations based on semantic or pixel similarity, suggesting that our co-occurrence frequency metric captures deeper concept connections in the hidden space of multimodal systems. Notably, this aligns with the Platonic Representation Hypothesis [Huh et al., 2024]: as multimodal systems scale, their internal representations tend to converge toward modeling the joint statistical structure of real-world events, despite differences in architecture, data, or training methods.

### 4.3 Hidden Language: the Connections Between Concepts

Our framework has excellent dynamic scalability. As test time compute scales, we gradually construct an increasingly comprehensive "world map" of the system's hidden language!

**Basic Connection** The map starts from our 150-concept vocabulary in Telescope. We visualize connections for 10 example concepts in Figure 5(a), where the color intensity reflects connection strength, highlighting better-trained pairs, and as it scales up, revealing the system's generalization progress: stronger generalization capability will lead to more uniform heatmap distributions. We also present some visualization results in Figure 4.

**Local and Global Connection** In our framework, each new telephone game tends to build new connections, linking existing and newly emerging concepts. As connections accumulate, genuine neighbors consistently reappear and occasional ones are submerged, gradually shaping a stable and

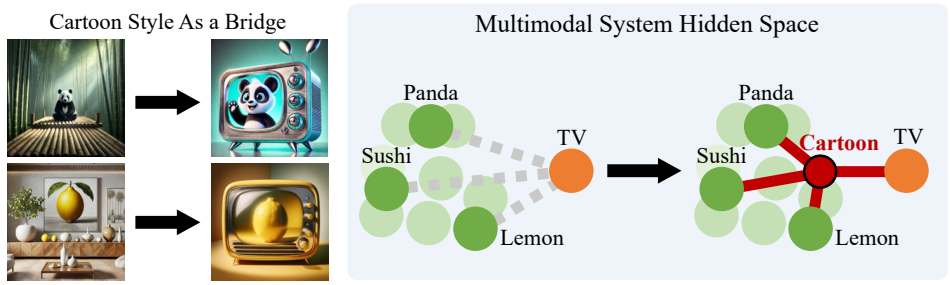



Figure 6: The intermediary node forms stable pathways between weakly connected concepts, making previously unstable combinations more reliably appear in generated images.

accurate local structure (Figure 5(b)). As the framework further expands, connections emerge between distinct local structures, gradually weaving a unified global network (Figure 5(c)). This integration bridges isolated local clusters and enriches the system by enabling cross-domain information exchange. In Appendix E, we provide a detailed explanation of the building details of such a graph structure in the "hidden language world map" of multimodal hidden space.

### 4.4 Complex Pattern and Concept Bridge

| Pattern | Crash Ratio |
|---|---|
| Van Gogh Style | 0.767 |
| Wood Texture | 0.560 |
| TV | 0.740 |
| TV (*improved*) | **0.427** |

Table 2: The crash ratio of our complex pattern and the "bridged improved" results on Pattern TV.

As detailed in Section 3.3, our Telescope includes 450 complex pattern concept pairs, derived from 150 common concepts across 3 patterns. We run the telephone game on these pairs using the OpenAI system. In Table 2, we report the concept crash ratios. Among the patterns, "Van Gogh Style Painting" (abbreviated as "Van Gogh Style") and "Displaying on a TV Screen" (abbreviated as "TV") show notably more fragile connections than simple-pattern in Section 4.2. For example, the system has less fitting to the scenario "displaying concepts on a TV screen" during the training phase. It shows that, after learning a limited number of these scenarios, the system has not developed robust generalization capabilities to extend this understanding to other concepts.

We explore bridging in the hidden space by introducing intermediary concepts, e.g., "Cartoon Style" or "Advertising Format", in Pattern TV. For each intermediary, we conduct experiments on 150 concepts to build a "TV"-centered hidden language local map like Figure 5(b). This map helps identify effective intermediary nodes for stabilizing fragile connections: using the concept"Cartoon Style", we form more stable pathways, as shown in Table 2 and Figure 6. As the map expands, we are committed to discovering more such bridges to enhance superalignment between multimodal system inputs and outputs.

## 5 Discussion

In this study, we quantify the concept connection strengths in multimodal systems' hidden space, also termed "hidden language", using our telephone game framework, the concept co-occurrence frequency metric, and the Telescope dataset. This hidden language reveals the bias from imbalanced training, tracks system generalization progress, and helps improve concept presentation in systems' output. We also use Reasoning-LLMs to infer how the multimodal systems' hidden language understands and simulates the world. Crucially, this is a test-time scalable framework: As the computation scales, an increasingly comprehensive multimodal hidden language world map will unfold in front of our eyes.

Due to large-scale systems' diverse output, our framework may be influenced by inherent randomness. In our future works, we will continuously conduct our telephone game to complete the hidden language world map, alleviating the impact of randomness. We will also advance on directed path formation to support tasks in specific domains. Additionally, we will apply various graph algorithms to this graph-based world map to identify optimal pathways between concepts.

## References


Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

StepFun. Introducing stepfun, March 2024. URL https://huggingface.co/stepfun-ai.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

OpenAI. Addendum to gpt-4o system card: Native image generation. *OpenAI technical report*, 2025.

Hila Chefer, Oran Lang, Mor Geva, Volodymyr Polosukhin, Assaf Shocher, Michal Irani, Inbar Mosseri, and Lior Wolf. The hidden language of diffusion models. *arXiv preprint arXiv:2306.00966*, 2023.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander Toshev, Jonathon Shlens, Ruoming Pang, and Yinfei Yang. Stair: Learning sparse text and image representation in grounded tokens. *arXiv preprint arXiv:2301.13081*, 2023.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

Juntu Zhao, Junyu Deng, Yixin Ye, Chongxuan Li, Zhijie Deng, and Dequan Wang. Lost in translation: Latent concept misalignment in text-to-image diffusion models. In *European Conference on Computer Vision*, pages 318–333. Springer, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

Tianhong Li, Qinyi Sun, Lijie Fan, and Kaiming He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Steven Derby, Paul Miller, Brian Murphy, and Barry Devereux. Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. *arXiv preprint arXiv:1809.02534*, 2018.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

OpenAI. Dall·e 3 system card. *OpenAI technical report*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, 58(347-352):240–242, 1895.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

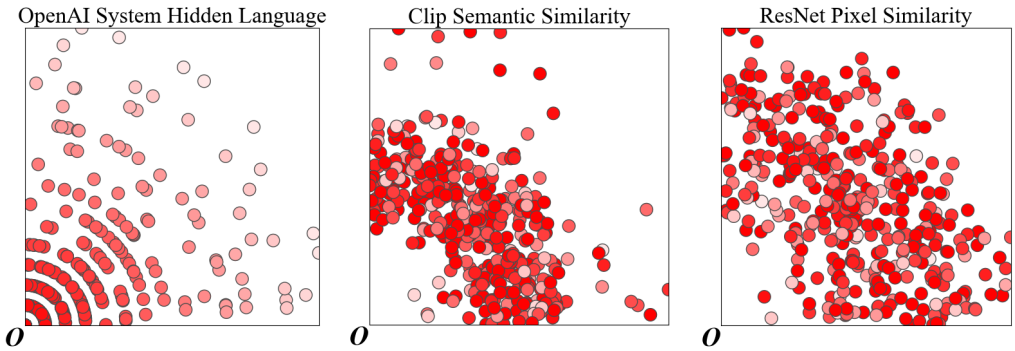
Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Edsger W Dijkstra. A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: his life, work, and legacy*, pages 287–290. 2022.

Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345–345, 1962.

Richard Bellman. On a routing problem. *Quarterly of applied mathematics*, 16(1):87–90, 1958.

# A Metric Correlation




(a) OpenAI System (b) Semantic Similarity (c) Visual Similarity

Figure 7: The correlation between the 3 metrics. Each point represents a concept pair. The distance from the origin, $O$, indicates their similarity under the current metric, and its color intensity reflects the strength of their connection in the hidden space of the multimodal system, i.e., OpenAI System.

In our Main Paper Section 4.2, we introduce the quantitative correlation between the co-occurrence frequency, semantic and visual similarity. Here, we present an intuitive scatter plot where each point represents the connection between two concepts. The distance from the origin point $O$ indicates their similarity under a given metric, while color intensity reflects their connection strength in the multimodal hidden space (i.e., co-occurrence frequency). For better visualization, concept pairs with the same similarity are randomly placed along the same-radius arc around the origin point $O$, rather than overlapping at a single point—resulting in a clearer 2D scatter plot. We observe that neither semantic nor visual similarity can adequately explain the concept connections in multimodal systems.

# B Visualization

Figure 8: The extended telephone game of the teaser example in the Main Paper Figure 1, and the results eventually stabilize with the ”dog in a frame.”

In this section, we will present the visualization results during the telephone game. First, in Figure 8, we extend the teaser example of GPT-4o-IG(20250325) shown in our Main Paper Figure 1 by continuing the telephone game. The results eventually stabilize with the “dog in a frame,” indicating that the connection between “frame” and “dog” is indeed stronger than that between “TV” and “dog”.

In Figure 9, we present additional concept pairs that remain stable during telephone game in our experiments, indicating strong concept connections in the multimodal hidden space, and Figure 10 shows more examples of concept pairs that exhibit concept crashes, indicating fragile concept connections in the multimodal hidden space.


Figure 9: Several visualization results of concept pairs that remain stable during telephone game in our experiments, indicating strong concept connections in the hidden space.


Figure 10: Several visualization results of concept pairs that exhibit concept crashes during telephone game in our experiments, indicating fragile concept connections in the hidden space.

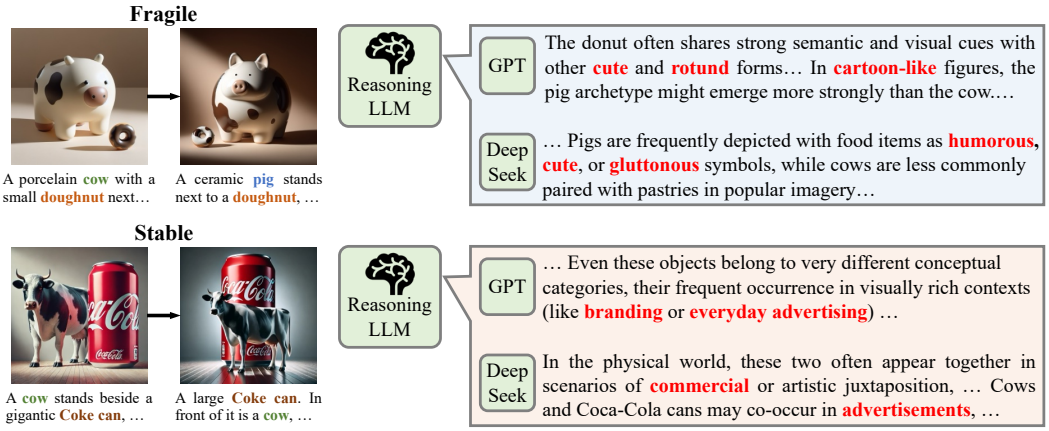



Figure 11: For hidden language beyond semantic or visual explanation, the LLMs Reasoning Analysis offers valuable insights into how multimodal systems understand the world.

## C LLMs as "MLPs"

We employ the Reasoning-LLMs and try to explore the system's hidden language about how it understands and simulates real-world physical laws. We use 2 SOTA Reasoning-LLMs: GPT-o1 [Jaech et al., 2024] and DeepSeek-R1 [Guo et al., 2025]. In Figure 11, we show 2 examples representing fragile and stable connections. Leveraging their reasoning capabilities, Reasoning-LLMs offer insights into the system's understanding of world patterns. For example, milk and coke often co-occur in beverage areas, and cows appear frequently on milk packaging, leading to a stable connection between cow and coke. These phenomena transcend semantic and visual similarities, revealing that multimodal systems are attempting to understand and simulate the common-sense laws in the human world. For the instruction prompt, please refer to Appendix G.

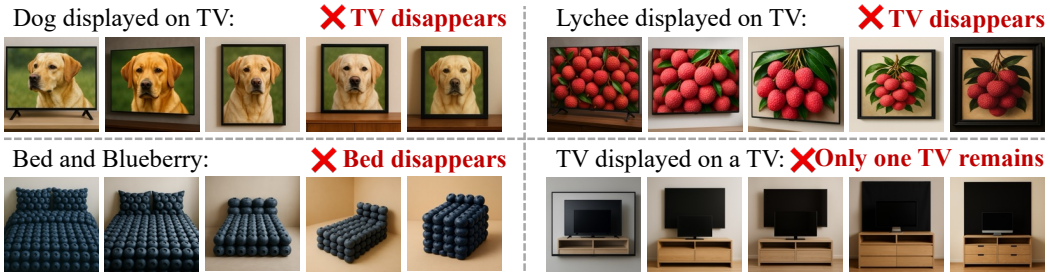



Figure 12: Visualization examples using the latest GPT-4o-IG(20250325). Despite its enhanced prompt-following capabilities, the system still reveals certain preference biases.

## D The Latest Released System

OpenAI has released the GPT-4o-IG(20250325), claiming improved prompt-following ability and deep integration with GPT-4o. This system is highly representative, transforming GPT-4o into a completely multimodal black-box system. In Figure 12, we conduct experiments on some representative examples using the web version. We observe that it still exhibits certain preference bias during the telephone game, which will aid our exploration of its hidden language.

## E Connection Graph Structure

As discussed in our Main Paper Section 4.3, we can construct an increasingly comprehensive "hidden language world map" of the multimodal system's hidden space by accumulating more and more telephone games.

In this graph structure, each node represents a concept, and the edges between nodes indicate the connection strength between concept pairs in the hidden space. These strengths are quantified using co-occurrence frequencies derived from a large number of repeated telephone games. Specifically, the co-occurrence frequency between two concepts, $A$ and $B$, is calculated only from telephone games where $A$ and $B$ are present in the initial input, or newly emerged together during the process.

We can apply various graph algorithms [Dijkstra, 2022, Floyd, 1962, Bellman, 1958] to this "world map" structured as a graph to find optimized paths between concepts, enhancing their connections to be more natural and stable in the multimodal system's output. For example, in our Main Paper Section 4.4, we observe that many concepts fail to appear consistently on a TV. By examining the local graph map centered around the concept "TV", we find that using the concept "cartoon style" as an intermediary node significantly improves the average connection strength. Depending on the number of additional elements we want to introduce, we can select multiple intermediary nodes and then apply graph algorithms such as the shortest path (the stronger the connection, the shorter the distance between 2 concepts) to identify strong bridges between concepts.

## F Open-Sourced Systems

We use features from open-source models to observe and validate that, during the process of compressing images into text and then reconstructing them into images, the system tends to favor concept combinations with stronger connections in its hidden language. We utilize the CLIP model [Radford et al., 2021], one of the fundamental components in open-source multimodal systems, to extract features from both images and text, which are then used to compute the similarities.

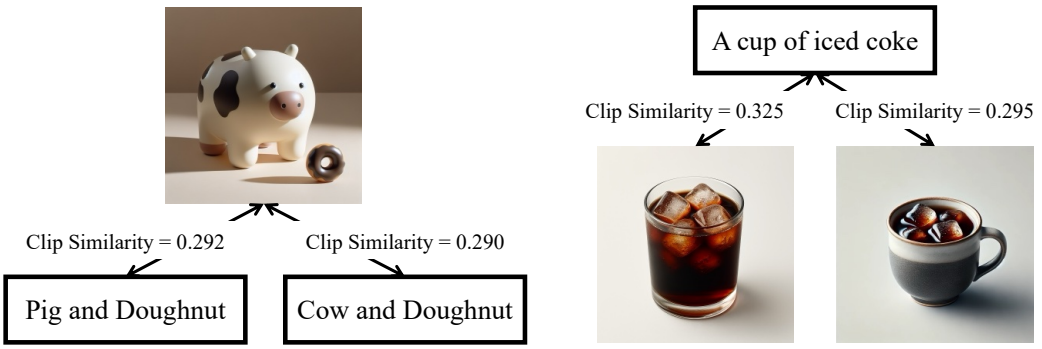



Figure 13: Preference bias of multimodal systems towards input concept combinations.

During the image-to-text compression process, concepts in the input image shift toward stronger concept combinations in the hidden space. For example, in Figure 13, the systems prefer to interpret the image of "cow and doughnut" as a feature more similar to "pig and doughnut," probably because pigs are more closely associated with doughnuts during training [Zhao et al., 2024].

During the text-to-image reconstruction process, concepts in input text shift toward stronger concept combinations in the hidden space. For example, in Figure 13, the systems prefer to interpret the text of "a cup of iced coke" as a feature more similar to the scene of a glass of iced coke, probably because glasses are more closely associated with iced coke during training [Zhao et al., 2024].

## G Instruction Prompt

Here we present the instruction prompt used in telephone game and Reasoning-LLMs.

### G.1 Get Start Description

As discussed in Section 3.1, LLMs convert concept pairs in our Telescope dataset into the start description for our telephone game. Here is the instruction prompt (To avoid interference from quotation marks in the system's response, we did not require it to output in JSON format.):

# Task
Execute the text-to-image prompt generation task under these protocols:

## Context Reset
- Two concepts the user provide are {A} and {B}
- The user wants to get a prompt describing these two concepts for creating an image

## Specifications
- If {A} and {B} are too different in SIZE, describe the scene as {A} and {B} are positioned next to each other. The background should be pure white without any other content.
- Else, describe the scene where {A} and {B} are most suitable when appearing. The background should be pure white without any other content.
- Refuse any detail or atmosphere description

## Compliance Requirements
- Strictly adhere to AI ethics guidelines
- Auto-filter NSFW/prohibited content
- Maintain objective observational stance

# Response Format
Only return your generated description, do not say anything else!

### G.2 Get Image Description

In telephone game, we need the multimodal systems to compress the images into text descriptions. Here is the instruction prompt (To avoid interference from quotation marks in the system's response, we did not require it to output in JSON format.):

# Task
Execute the visual description task under these protocols:

## Description Specifications - Generate single-paragraph visual depiction
- Apply 'object + attributes + spatial relationships' framework
- Include comprehensive detailed visual elements:
  – All of the details in the image
  – Chromatic composition & material textures
  – Lighting dynamics & spatial perspective
  – Stylistic signatures & atmosphere creation

## Compliance Requirements
- Strictly adhere to AI ethics guidelines
- Auto-filter NSFW/prohibited content
- Maintain objective observational stance

# Response Format
- Begin directly with the description, no phrases like 'The image shows or describes or presents'
- Only return your description, do not say anything else!

### G.3 Check Concept Existence

In the co-occurrence frequency metric, we use LLMs to check if a concept exists in the image description. Here is the instruction prompt:

# Task
Execute the Concept Existence Identify task under these protocols:

## Context Reset
- Focus exclusively on the currently provided description
- You should check the existence of the concept, and check if new concepts appear in the description, instead of the background
- Provided description: '{description}'
- Concepts I want to check: '{concept list}'

## Note
- You should first check if the concepts I want to check :'{concept list}' appear in the description scene, using True or False to indicate (If a concept I want to check is described by an alias, we consider it to appear)
- If there are new OBVIOUS concepts (not include background), add a new concept in return and use True to indicate
- Before adding any new concept, you must check it first:
  – SUBspecies: Any concept that are of the same species or subspecies as existing concepts: '{concept list}' should NOT be considered as a new concept. For example, cat and dog are different, but cat and ragdoll cat are the same. Cow and horse are different, but cow and bull are the same. Turtle and tortoise are the same.
  – Breed: If the new concept is a breed name of an existing concept: '{concept list}', it should NOT be considered as a new concept. For example, golden retriever is not a new concept (because it is the same as dog).
  – Different Term: Different words used to represent different ages of existing concepts: '{concept list}', it should NOT be considered as a new concept. For example, kitten is not a new concept (because it is the same as cat), and puppy is not a new concept (because it is the same as dog).
  – Background: The background and environment should NOT be considered as a new concept.
  – Light: The light of should NOT be considered as a new concept.
  – Part: The part of an existing concept's body, should NOT be considered as a new concept.
  – Style: The style/sense/feeling of the whole image should NOT be considered as a new concept.
  – Texture: The texture of an existing concept, should NOT be considered as a new concept.
  – Representation: The representation of an existing concept, such as a painting, a sculpture, a toy,etc., should NOT be considered as a new concept.

# Response Format
- Only return a json, do not say anything else!
- The json format:
```
{
  "{ori_concept}": your decision (True or Fasle)
}
```
- If you find a new concept, you must check if it is a new concept. And if it is a new concept, add it to the return json.

### G.4 Reasoning LLMs

We use Reasoning-LLMs to get insights into how multimodal systems understand the world. Taking the example of “Cow and Coke”, here is the instruction prompt:

# Task
Execute the Concepts Potential Connections Reasoning task under these protocols:

## Background
- I am experimenting with a multimodal system that uses the ”Pre-Description” to generate an image and then uses text to describe the image to get the ”New Description”.
- We find that in the above image reconstruction process, due to the system’s existing preferences, it tends to reconstruct the concept combination that is more closely related in its understanding, which we call the Hidden Language of multimodal systems.
- The ”Pre-Description”: A majestic black and white Holstein cow stands confidently beside a gigantic red Coca-Cola can, casting distinct shadows on the hardwood floor below. The cow’s coat boasts a clean sheen, highlighting its stark contrasts between patches of black and white, while its expressive eyes and symmetrical horns add to its regal presence. The Coca-Cola can towers over the cow, flaunting a vibrant red hue that dominates the scene with its glossy metallic texture reflecting bright lighting. The iconic white script logo of Coca-Cola is embossed on the can’s surface, creating a strong branding visual impact. The scene features dynamic lighting that illuminates details on both subjects, generating subtle reflections and emphasizing seamless spatial coexistence in the serene, minimalist atmosphere.
- The ”New Description”: A large Coca-Cola can, predominantly bright red with the signature white logo curving across its shiny metallic surface, stands upright as an imposing backdrop. In front of this towering can is a black and white cow, which appears proportionally smaller, painted with a realistic sheen on its smooth, glossy skin, suggesting a polished, almost plasticine texture. The cow’s body casts a long, soft shadow against the reflective hardwood floor, suggesting a light source positioned above and slightly in front of the duo, which faintly illuminates the details such as the cow’s textured coat and the can’s metallic glint. The setting exudes a surreal, hyperrealistic atmosphere, achieved through the juxtaposition of everyday elements at contrasting scales, creating a unique blend of reality and artistic abstraction.

## Experimental Findings
- I find that two seemingly unrelated concepts, ”Cow” and ”Coke Can”, are retained stably.

## My Needs
- Please infer and analyze the reasons why these two seemingly unrelated concepts are closely related from the perspective of the multimodal system hidden space. What laws in the physical world it may reflect.

# Response Format
```
{
   ”Reason”: Describe your reasons.
}
```