Emergent Symbol Grounding in Language Models

Anonymous Author(s)

Affiliation

Address

email

1 Introduction

Recently, several studies have described "grounding" as an emergent property of vision-language models [VLMs; 15, 1, 6, *inter alia*] trained without explicit grounding objectives. Evidence comes from two main observations: spatial localization, where attention patterns and prompt-conditioned activations align tokens with fine-grained image regions or object boundaries [5, 4]; and cross-modal geometric correspondences that can be recovered even without paired image–text supervision [18]. In this context, "emergence" simply means that such alignment patterns arise without explicit, fine-grained supervision, and "grounding" denotes statistical associations learned from large-scale co-occurrence of pixel regions and phrases in captions.

This behavioral perspective departs from the original formulation of the symbol grounding problem [11], which emphasizes how symbols acquire referentially anchored, causally supported links to percepts and actions. Such links should remain stable across contexts and be informative under interventions, rather than merely arising as byproducts of distributional regularities. Siskind [19] demonstrated with an algorithmic model that word-meaning mappings can be learned without relying solely on high co-occurrence correlations, underscoring the need to test grounding with mechanisms beyond surface statistics. To address the fundamental research question of how, if at all, symbol grounding emerges in autoregressive language modeling, we must move beyond correlations and probe causal mechanisms that link symbols to their environments, while also examining training dynamics to identify when and how such links appear.

2 Method

2.1 Dataset and Tokenization

Our guiding principle is that symbol grounding should be learned from multimodal interactions, where environmental context provides the ground and dialogue supplies the linguistic forms. To capture this, we design a custom word-level tokenizer in which every lexical item is represented in two forms: one token that appears in non-verbal descriptions (e.g., a *box* in the scene) and another that appears in utterances (e.g., *box* in speech). We call these environmental tokens (<ENV>) and linguistic tokens (<LAN>). For example, box$_{\texttt{<ENV>}}$ and box$_{\texttt{<LAN>}}$ are treated as entirely separate tokens, so any successful mapping must be learned during training rather than inherited from token identity. We instantiate this framework in three datasets, ranging from child-directed transcripts to image-based dialogue. **Due to space limits, we show CHILDES results in the main paper and leave others to the Appendix.**

Grounded Child-Directed Speech. The Child Language Data Exchange System (CHILDES) corpus [14] provides transcripts enriched with environmental annotations. We preprocess the data by retaining only two streams: (i) spoken utterances, mapped to linguistic tokens (<LAN>), and (ii) environmental descriptions, mapped to environmental tokens (<ENV>). Environmental context is drawn from three annotation types:

- **Local Events**: simple events, pauses, long events, or remarks interleaved with transcripts.
- **Action Tiers**: actions performed by the speaker or listener (e.g., `%act: runs to toy box`). These also include cases where an action replaces speech (e.g., `0 [% kicks the ball]`).

Submitted to 1st Workshop on Interpreting Cognition in Deep Learning Models (CogInterp @ NeurIPS 2025).

- **Situational Tiers**: situational information tied to utterances or to larger contexts (e.g., `%sit: dog is barking`).

Caption-Grounded Dialogue. The Visual Dialog dataset [8] pairs MSCOCO images [12] with multi-turn dialogues collected through sequential question–answer exchanges about each image. In our setup, the MSCOCO captions serve as the environmental stream (environmental tokens <ENV>), while the dialogue turns form the linguistic stream (linguistic tokens <LAN>). This design provides a controlled multimodal setting in which textual descriptions of visual scenes grounds natural conversational interaction. Compared to CHILDES, this setup introduces richer semantics and longer utterances, while still using the same text-only inputs for both streams, thereby offering a stepping stone toward grounding in fully visual contexts.

Image-Grounded Dialogue. To move beyond textual proxies for the environment, we construct a fully image-grounded dialogue setup. Here, the environmental stream is derived directly from images: each image is tokenized into patch embeddings using a frozen vision transformer (ViT, [9]), and each patch embedding is treated as an <ENV> token, analogous to the visual tokens in modern VLMs. We adopt DINOv2 [16] as our visual encoder, since it is trained purely from vision without auxiliary text supervision (in contrast to CLIP [17]), ensuring that environmental tokens reflect visual structure alone. Dialogue remains in <LAN> form, yielding a realistic multimodal interaction where conversational predictions are conditioned directly on visual input. This setup allows us to evaluate whether the grounding effects observed in controlled text-based contexts also extend to real multimodal settings.

2.2 Evaluation Protocol

Behaviorally, we assess symbol grounding with a contrastive test that asks whether a model assigns higher probability to the correct linguistic token when the matching environmental token is present, similar to priming in psycholinguistics. This evaluation applies uniformly across datasets: in CHILDES and caption-grounded dialogue, environmental priming comes descriptive templates; and in image-grounded dialogue, from ViT-derived visual tokens. We compare the following conditions:

- **Match (Experimental Condition)**: The environmental context contains the correct <ENV> token for the target word, and the model is expected to predict its <LAN> counterpart.
- **Mismatch (Control Condition)**: The environmental context is replaced with an <ENV> token from a different vocabulary item (sampled from the remaining words). The model is still tasked with predicting the same <LAN> token, and surprisal is measured, but performance is expected to be at chance level given the absence of a matching environmental cue.

For example, when evaluating the word $\texttt{box}_{\texttt{<LAN>}}$, the input sequence is: <CHI> $\texttt{jumped}_{\texttt{<ENV>}}$ $\texttt{beside}_{\texttt{<ENV>}}$ $\texttt{a}_{\texttt{<ENV>}}$ $\texttt{large}_{\texttt{<ENV>}}$ $\texttt{box}_{\texttt{<ENV>}}$ <CHI> $\texttt{I}_{\texttt{<LAN>}}$ $\texttt{found}_{\texttt{<LAN>}}$ $\texttt{a}_{\texttt{<LAN>}}$ $\texttt{cool}_{\texttt{<LAN>}}$ ____, where the model is expected to predict $\texttt{box}_{\texttt{<LAN>}}$. In the control (mismatch) condition, the environmental token $\texttt{box}_{\texttt{<ENV>}}$ is replaced by another valid noun such as $\texttt{chair}_{\texttt{<ENV>}}$. Formally, let θ denote model parameters. For a target type v with linguistic token v_{LAN} and environmental token v_{ENV}, let $c \in C_v$ be a context template and let $u \neq v$ index a mismatched type. In the match condition, $e = \text{match}(v)$ inserts $v_{\texttt{<ENV>}}$ into the environmental stream of c. In the mismatch condition, $e = \text{mismatch}(u)$ inserts $u_{\texttt{<ENV>}}$ with $u \neq v$ while the target remains $v_{\texttt{<LAN>}}$.

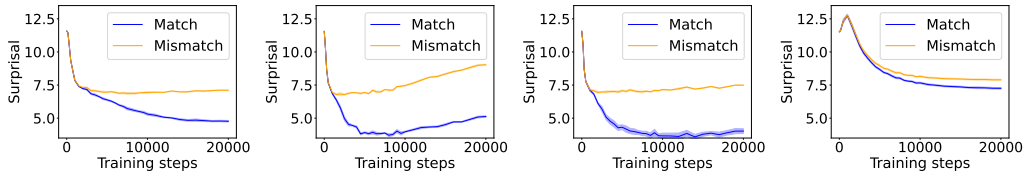
Grounding Information Gain. Following prior work, we quantify prediction quality using the mean surprisal of a word. For a token w with context c and environmental ground e, surprisal is defined as $s_\theta(w \mid c, e) = -\log_2 P_\theta(w \mid c, e)$, which quantifies the unexpectedness (self-information) of predicting w. In our setting, $w = v_{\texttt{<LAN>}}$ and e specifies either the correct ground $v_{\texttt{<ENV>}}$ (match) or a mismatched ground $u_{\texttt{<ENV>}}$ with $u \neq v$.

For each target word v, we construct N template contexts $\{c_1, \ldots, c_N\}$. For each template c_n, we evaluate the model under the match condition and under M mismatched conditions. The *Grounding Information Gain* $G(v)$ for v is

$$G(v) = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{1}{M} \sum_{u \neq v}^{M} \left[s_\theta(v_{\texttt{<LAN>}} \mid c_n, u_{\texttt{<ENV>}}) - s_\theta(v_{\texttt{<LAN>}} \mid c_n, v_{\texttt{<ENV>}}) \right] \right). \tag{1}$$

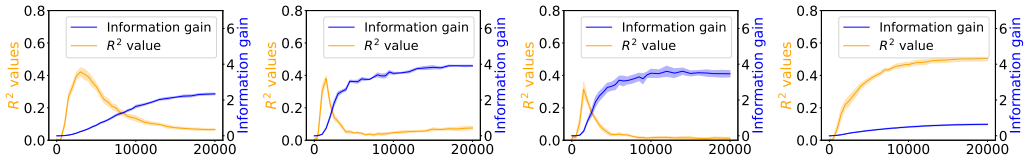
This is an estimation (via sampling) of,

$$G(v) = \mathbb{E}_{c,u} \left[\log_2 \frac{P_\theta(v_{\texttt{<LAN>}} \mid c, v_{\texttt{<ENV>}})}{P_\theta(v_{\texttt{<LAN>}} \mid c, u_{\texttt{<ENV>}})} \right]. \tag{2}$$



(a) 12-Layer Transformer. (b) 12-Layer Mamba 2. (c) 4-Layer Mamba 2. (d) 4-Layer LSTM.

Figure 1: Average surprisal of the experimental and control conditions over training steps.



(a) 12-Layer Transformer. (b) 12-Layer Mamba 2. (c) 4-Layer Mamba 2. (d) 4-Layer LSTM.

Figure 2: Grounding information gain and its correlation to the co-occurrence of linguistic and environment tokens over training steps.

Intuitively, $G(v)$ is the expected log-likelihood ratio between the match and mismatch conditions. It quantifies how much evidence the correct environmental ground provides for the hypothesis that the observed linguistic form was generated from it, or equivalently, the average number of bits by which uncertainty is reduced when conditioning on the correct ground rather than an incorrect one. A positive value of $G(v)$ indicates that the correct environmental ground increases the predictability of its linguistic form, measured in bits. At the dataset level, we report $G_\theta = \frac{1}{|V|}\sum_{v\in V} G(v)$, and track $G_\theta^{(t)}$ across training steps t to analyze how grounding emerges over time.

Tested Vocabulary and Test Context. We select a vocabulary of 100 nouns from the MacArthur–Bates Communicative Development Inventories (CDI) [10] that occur frequently in our training corpus. Each word serves once as the target, with the remaining $M = 99$ nouns used to construct mismatched conditions. For each noun mentioned above, we create $N = 10$ different contexts. Each context has two parts: the first part describe an environment that a child is in (with <ENV> tags), and the second part is a sentence they say (with <LAN> tag). The two parts both contain the target noun. The full word list and context templates are provided in the Appendix.

2.3 Model Training

We train autoregressive language models from random initialization, ensuring that no prior linguistic knowledge influences the results. Training uses the standard causal language modeling objective, as in most generative LMs. To account for variability, we repeat all experiments with 5 random seeds, randomizing both model initialization and corpus shuffle order. Our primary architecture and training setups follow a Transformer [20] in the style of GPT-2, though we also run parallel experiments with state-space models [7] following Mamba-2. Standard LLaVA [13] uses a two-layer perceptron to project ViT embeddings into the language model. In our case, since DINOv2 provides embeddings with the same dimensionality as the language model, we bypass this projection and feed the representations directly into the autoregressive model. We obtain the developmental trajectory of the model by saving 33 checkpoints at various training steps.

3 Experiments

3.1 Behavioral Evidence of Emergent Grounding

In this section, we ask: **Does symbol grounding emerge behaviorally in autoregressive language models?** For Transformers and Mamba-2, surprisal in the match condition decreases steadily while the mismatch condition plateaus higher, indicating that the model exploits the environmental ground to predict the linguistic form. The 12-layer version shows sharper learning dynamics initially and drifts later, consistent with the tendency of Mamba to overfit local pattern shortcuts [22]. In contrast, the unidirectional LSTM (Figure 1d) shows little separation between the conditions, reflecting sequential state compression and the absence of content-addressable retrieval. Overall, these results provide behavioral evidence of emergent grounding: in sufficiently expressive architectures, the correct environmental ground reliably lowers surprisal for its linguistic counterpart, whereas LSTMs fail to exhibit this effect, marking an architectural boundary on where grounding can emerge.

3.2 Behavioral Effects Surpass Co-occurrence

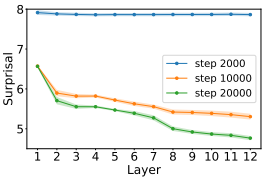
A natural concern is that surprisal reductions might reflect shallow statistics: **the model could simply memorize frequent co-occurrences of <ENV> and <LAN> tokens rather than learn a general mapping.** To test this, we compare each word's co-occurrence frequency with its Grounding Information Gain. Raw counts are log-transformed to match surprisal's log scale, and for every checkpoint we regress log co-occurrence counts against Grounding Information Gain, yielding an R^2 over training. Figure 2 plots R^2 (orange) alongside overall gain (blue). In both Transformer and Mamba-2, R^2 rises early but soon collapses, while Grounding Information Gain continues to grow. By contrast, the LSTM shows persistently high R^2 but negligible gain. Thus, in Transformers and Mamba-2, grounding cannot be explained by co-occurrence alone: models initially exploit surface regularities, but later improvements depend on richer internal mechanisms.

3.3 An Mechanistic Interpretability Account of Grounding

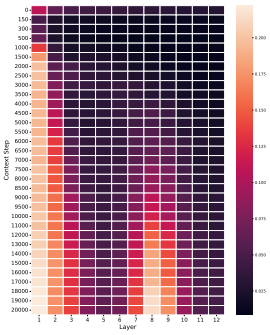
To give a mechanistic account of symbol grounding, e.g., where and how in the network is it represented, we apply two interpretability analysis.

Tuned Lens Probing. We probe layer-wise representations using tuned lens [2], which trains affine translators to map intermediate activations to the final prediction space while keeping the output head fixed. Figure 3a shows surprisal estimates from each layer under the match condition. Early layers remain poor predictors, while surprisal drops substantially beginning at layer 7, pointing to a representational shift in the mid layers.

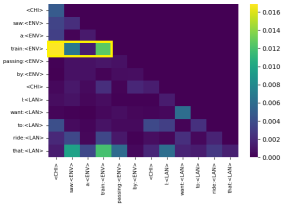
Saliency Flow. For each attention layer l, we compute a saliency matrix following Wang et al. [21]: $I_l = \left|\sum_h A_{h,l} \odot \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}}\right|$, which measures the contribution of attention heads (via gradient $\times$ attention) to the loss. We focus specifically on anchor-to-end connections, i.e., flows from environmental ground (<ENV>) to their linguistic form (<LAN>). Figure 3b shows that anchor-to-end saliency is weak in early layers but rises sharply in later training steps, peaking in layers 7–9. This suggests that mid-layer attention plays a central role in establishing symbol–ground correspondences.



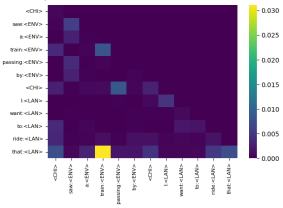
(a) Layer-wise tuned lens to predict the match condition.



(b) Saliency of layer-wise attention from environmental to linguistic tokens across training steps.



(c) A gather head identified in layer 6 head 7.



(d) A aggregate head identified in layer 7 head 12.

Figure 3: Mechanistic analysis of symbol grounding emergence.

3.4 Attention-Head Mechanisms Implements Symbol Grounding

We next ask how this computation is implemented at the level of individual attention heads. Bick et al. [3] has argued that autoregressive LMs employ a Gather-and-Aggregate (G&A) mechanism: some heads compress relevant information into a subset of positions (gather), while others redistribute this information to downstream tokens (aggregate).

Figure 3c shows an example gather head (layer 6, head 7). Here, attention concentrates on environmental ground tokens such as $\texttt{train}_{\texttt{<ENV>}}$, pooling information from previous tokens into a compact representation. This behavior suggests that early mid-layer heads are responsible for collecting grounded information into retrievable states. Figure 3d highlights an aggregate head (layer 7, head 12). In contrast to the gather pattern, this head propagates information from the environmental ground ($\texttt{train}_{\texttt{<ENV>}}$) to the corresponding linguistic tokens ($\texttt{train}_{\texttt{<LAN>}}$ and related context). This pattern reflects a broadcasting role, where collected grounding information is distributed to the locations that drive final predictions. Taken together with our earlier identification of gather heads in lower layers (prior to layer 7), these results suggest that symbol grounding is implemented via a functional specialization across attention heads. Specifically, gather heads in early layers compress and store information from environmental grounds into a smaller set of hidden positions, creating retrievable intermediate representations. Aggregate heads in mid layers (particularly layers 7–9) propagate this stored information forward, redistributing it to the positions associated with linguistic tokens and thereby supporting accurate prediction.

References

[1] Anthropic. The claude 3 model family: Opus, sonnet, haiku, March 2024. URL https://www.anthropic.com/news/claude-3-family.

[2] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[3] Aviv Bick, Eric P. Xing, and Albert Gu. Understanding the skill gap in recurrent models: The role of the gather-and-aggregate mechanism. In *Forty-second International Conference on Machine Learning*, 2025.

[4] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024.

[5] Shengcao Cao, Liang-Yan Gui, and Yu-Xiong Wang. Emerging pixel grounding in large multimodal models without grounding supervision. *arXiv preprint arXiv:2410.08209*, 2024.

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[7] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071. PMLR, 2024.

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[10] Larry Fenson, Virginia A Marchman, Donna J Thal, Phillip S Dale, J Steven Reznick, and Elizabeth Bates. Macarthur-bates communicative development inventories. *PsycTESTS Dataset*, 2006.

[11] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3): 335–346, 1990.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[14] Brian MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press, 2014.

[15] OpenAI. Hello gpt-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024.

[17] 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.

[18] Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! towards vision-language correspondence without parallel data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24983–24992, 2025.

[19] Jeffrey Mark Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91, 1996.

[20] 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.

[21] Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, 2023.

[22] WangJie You, Zecheng Tang, Juntao Li, Lili Yao, and Min Zhang. Revealing and mitigating the local pattern shortcuts of mamba. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12156–12178, 2025.

A Appendix

Context Templates. All contexts are created with gpt-4o-mini, with the following prompt:

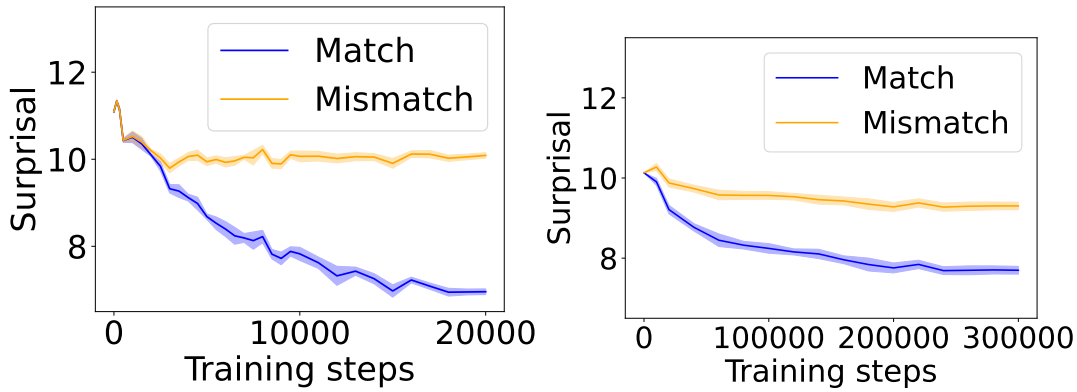
```
template = f'''given the word "{word}", create 3 pairs of sentence
that follow this requirement:
1. The first sentence has subject "The child", describing an event or situation,
and have the word "{word}". Make sure to add newline to the end of
this first sentence
2. The second sentence is said by the
child (only include the speech itself,
don't include "the child say", etc.),
and the word "{word}" also appears in the sentence
said by the child. Do not add quote marks either
3. Print each sentence on one line.
Do not include anything else.
4. Each sentence should be short, less than 10 words.
5. The word "{word}" in both sentence have
the same meaning and have clear indication
or implication relationship.
6. "{word}" should not appear at the
first/second word of each sentence.
Generate 3 pairs of such sentences,
so there should be 6 lines in total.
You should not add number.
For each line, just print out the sentence.
'''
```

Word List for CHILDES and Vision Dialogue (Text Only). ['box', 'book', 'ball', 'hand', 'paper', 'table', 'toy', 'head', 'car', 'chair', 'room', 'picture', 'doll', 'cup', 'towel', 'door', 'mouth', 'camera', 'duck', 'face', 'truck', 'bottle', 'puzzle', 'bird', 'tape', 'finger', 'bucket', 'block', 'stick', 'elephant', 'hat', 'bed', 'arm', 'dog', 'kitchen', 'spoon', 'hair', 'blanket', 'horse', 'tray', 'train', 'cow', 'foot', 'couch', 'necklace', 'cookie', 'plate', 'telephone', 'window', 'brush', 'ear', 'pig', 'purse', 'hammer', 'cat', 'shoulder', 'garage', 'button', 'monkey', 'pencil', 'shoe', 'drawer', 'leg', 'bear', 'milk', 'egg', 'bowl', 'juice', 'ladder', 'basket', 'coffee', 'bus', 'food', 'apple', 'bench', 'sheep', 'airplane', 'comb', 'bread', 'eye', 'animal', 'knee', 'shirt', 'cracker', 'glass', 'light', 'game', 'cheese', 'sofa', 'giraffe', 'turtle', 'stove', 'clock', 'star', 'refrigerator', 'banana', 'napkin', 'bunny', 'farm', 'money']

Word List for Vision Dialogue (VLM). ["box", "book", "table", "toy", "car", "chair", "doll", "door", "camera", "duck", "truck", "bottle", "bird", "elephant", "hat", "bed", "dog", "spoon", "horse", "train", "couch", "necklace", "cookie", "plate", "telephone", "window", "pig", "cat", "monkey", "drawer", "bear", "milk", "egg", "bowl", "juice", "ladder", "bus", "food", "apple", "sheep", "bread", "animal", "shirt", "cheese", "giraffe", "clock", "refrigerator", "accordion", "aircraft", "alpaca", "ambulance", "ant", "antelope", "backpack", "bagel", "balloon", "barrel", "bathtub", "beard", "bee", "beer", "beetle", "bicycle", "bidet", "billboard", "boat", "bookcase", "boot", "boy", "broccoli", "building", "bull", "burrito", "bust", "butterfly", "cabbage", "cabinetry", "cake", "camel", "canary", "candle", "candy", "cannon", "canoe", "carrot", "cart", "castle", "caterpillar", "cattle", "cello", "cheetah", "chicken", "chopsticks", "closet", "clothing", "coat", "cocktail", "coffeemaker", "coin", "cosmetics"]

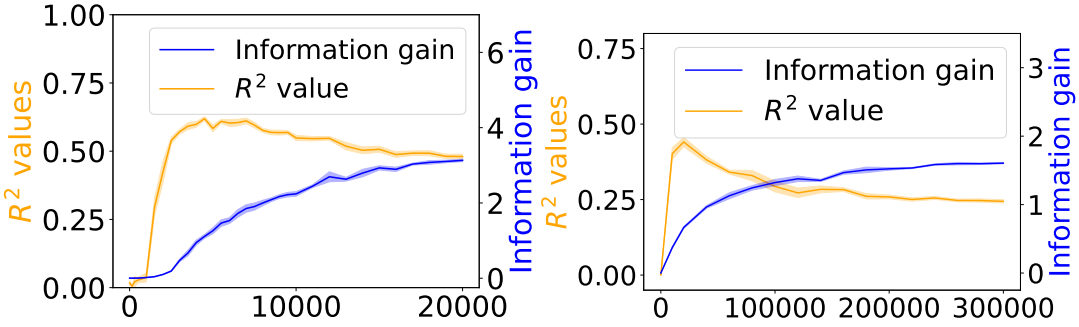
Checkpointing. We save the intermediate steps: [0, 150, 300, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, 6500, 7000, 7500, 8000, 8500, 9000, 9500, 10000, 11000, 12000, 13000, 14000, 15000, 16000, 17000, 18000, 19000, 20000] (33 checkpoints in total)

Visual Dialogue Results. We provide the Visual Dialogue results in Figure 4 and Figure 5.



(a) Vision Dialogue (Text Only) (b) Vision Dialogue (VLM)

Figure 4: Average surprisal of the experimental and control conditions over training steps.



(a) Vision Dialogue (Text Only) (b) Vision Dialogue (VLM)

Figure 5: Grounding information gain and its correlation to the co-occurrence of linguistic and environment tokens (or images) over training steps.