# LEARNING WORDS BY SIMULATING WORLDS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Result verbs such as *open*, *slice*, and *fill* describe not only what is present in a scene, but how an action transforms one world state into another. We study whether this causal structure provides a privileged learning signal for language models. We train an autoregressive model on interleaved language and visual transition sequences, where each example contains a triple of a pre-condition image, an action description, and a post-condition image. The model is trained with complementary objectives: inferring the action from an observed pre/post transition, predicting the post-condition from the pre-condition and action, and reconstructing the pre-condition from the post-condition and action. Across controlled ablations, visual transition simulation improves grounded word acquisition when it is paired with sufficient language supervision. The gains are concentrated on verbs rather than nouns, and within verbs, on result verbs rather than manner verbs, consistent with the fact that result verbs are defined by observable state changes while manner verbs depend more on motion trajectories. These findings suggest that grounding verb semantics requires more than aligning words with percepts: it requires predictive models of how the world changes.

## 1 INTRODUCTION

Children cannot understand and learn the verb *open* from a single image, as the word denotes a transition over time: something is closed, an action occurs, and the same thing becomes open. Many concrete verbs have this structure: they do not merely name what is visible in a scene; instead, they describe how agents, objects, and states are related through changes. To learn such words, a learner must perceive not only objects, but also the causal transformations in which those objects participate.

This observation raises a question for grounded language acquisition. Most benchmarks and methods for visual grounding treat language as a mapping from words to percepts: nouns to objects (Xu & Tenenbaum, 2007), adjectives to attributes (Buettner & Kovashka, 2024), and referring expressions to regions (Kazemzadeh et al., 2014; Yu et al., 2016). This formulation has been highly effective for object-centered grounding, but it is incomplete for actions (Ma et al., 2023). Verbs such as *open*, *slice*, or *fill* are not grounded in a single region or frame. Their meanings are distributed across an initial state (pre-condition) and a resulting state (post-condition). The perceptual evidence is not only what is present, but what has changed.

This distinction also exposes a limitation in contemporary Vision-Language Models (VLMs; Google, 2023; OpenAI, 2024; Anthropic, 2024, *inter alia*). Although VLMs can describe objects, actions, and events, they are still largely grounded through recognition, image-caption alignment, and instruction following: perception is used to answer prompts, not to model how the world changes (Gao et al., 2025). This is insufficient as grounding an action verb requires representing the causal transition it denotes: the state before the action, the state after it, and the relation by which one becomes the other (Gao et al., 2016). Such transition-centered representations are the basis of *world models*, which support prediction, monitoring, and planning by simulating how actions transform states (Wooldridge & Jennings, 1995; Ha & Schmidhuber, 2018; LeCun, 2022). From this perspective, current VLM failures on spatiotemporal perception, intuitive physics, future-state prediction, and agentic causality are not merely benchmark errors, but failures of grounded world modeling (Goyal et al., 2020; Shangguan et al., 2024; Fu et al., 2024; Zhang et al., 2025; Bakhtin et al., 2019; Yi et al., 2020; Bear et al., 2022). They motivate recent calls to move beyond reactive perception and brute-force long context toward models that construct, update, and predict with implicit world models from continual sensory experience (Yang et al., 2025).



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

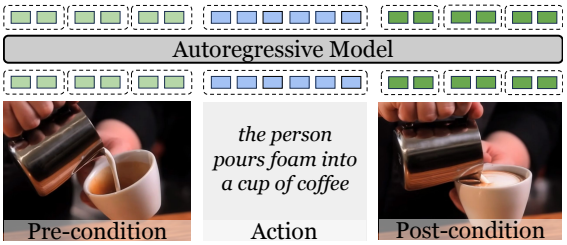



Figure 1: Overview of our autoregressive world-language model, illustrated with **forward simulation**.

Table 1: Data statistics for the real-video and world-model-distilled sources. Seen and unseen are curated subsets.

| Source | Split | Triples | Verbs | Nouns |
|---|---|---|---|---|
| Real-Video | Train | 351,044 | 400 | 1,016 |
| | Test | 3,999 | 80 | 407 |
| World-Model Distilled | Train | 488,703 | 1,788 | 5,322 |
| | Test (seen) | 47,332 | 97 | 1,220 |
| | Test (unseen) | 12,097 | 120 | 794 |

We see this as a grounded word acquisition problem. Our hypothesis is specific: for concrete predicates whose meanings are partly defined by observable state change, action-conditioned visual prediction should provide a privileged learning signal. If a verb is individuated by the transition it licenses, then predicting the resulting state from the initial state and the action should expose the learner to the variables that matter for that verb: the participants, their initial configuration, their resulting configuration, and the causal relation between them. This hypothesis does not predict that visual prediction should uniformly improve all words. It predicts a structured dissociation: *transition prediction should benefit verbs more than nouns, with the strongest gains for result verbs, whose meanings are defined by the states they bring about, rather than manner verbs, which depend more on how an action unfolds over time* (Levin, 2008).

We test this hypothesis in a modern autoregressive modeling setup, training on interleaved language and visual transition sequences to instantiate a joint autoregressive world-language model. Each example contains a pre-condition image, an action description, and a post-condition image. The model is trained with three complementary objectives: transition captioning (inferring the action from the pre-condition and post-condition), forward simulation (predicting the post-condition from the pre-condition and action), and backward simulation (inferring the pre-condition from the post-condition and action). These objectives place language prediction and visual state prediction in the same representational system, allowing us to ask whether simulation-like training changes the acquisition dynamics of concrete words.

Our results support this hypothesis. Action-conditioned simulation improves lexical acquisition, but only when paired with sufficient language-side supervision. Its benefits are not uniform: they are concentrated on verbs rather than nouns, and within verbs, on result verbs rather than manner verbs. This dissociation reflects the structure of the world-modeling signal: before-and-after images make changes of state directly observable, providing evidence for how actions transform the world, but only weakly encode how an action unfolds over time. Accordingly, simulation helps most when a word's meaning is defined by the outcome it brings about. These findings suggest that grounding is not merely mapping words to percepts; for verbs, it requires joint world-language modeling, where linguistic forms are grounded to the predictive representations of how actions transform the world.

## 2 Learning Language by Simulation

We minimally formulate neural word learning from visual simulation as a single autoregressive sequence model (Figure 1). Each training example is represented as an interleaved sequence of visual and textual units: a pre-condition image, an action description, and a post-condition image. A standard causal Transformer processes the sequence from left to right and is trained to predict the next element at each position. For text positions, the model uses the usual next-token prediction objective. For visual positions, instead of predicting pixels directly, the model predicts the next visual embedding, allowing visual states to be modeled in the same autoregressive framework as language.

### 2.1 Model Architecture and Implementations

We adopt the assumption that word learning can rely on a visual prior but not on a language prior, such as a pretrained language model. Accordingly, we initialize the visual backbone from NEPA (Xu et al., 2025), an autoregressive visual model pretrained with causal patch-embedding prediction, while initializing the language-specific components from scratch. We use whole-word tokenization

Table 2: Sequence formats and their roles. $I_{\text{pre}}$: pre-condition image. $I_{\text{post}}$: post-condition image. $a$: action text. Components are ordered from left to right. Underlining marks the response block under completion-only supervision, not the full-sequence loss mask.

| Format | Sequence order | Description |
|---|---|---|
| $\text{Sim}_{\text{fwd}}$ | $I_{\text{pre}},\ a,\ \underline{I_{\text{post}}}$ | Predicts the resulting visual state from the initial state and action description. |
| $\text{Sim}_{\text{bwd}}$ | $I_{\text{post}},\ a,\ \underline{I_{\text{pre}}}$ | Infers a compatible prior visual state from the observed result and action description. |
| $\text{LM}_{\text{fwd}}$ | $I_{\text{pre}},\ I_{\text{post}},\ \underline{a}$ | Predicts the action text associated with the observed pre/post transition, corresponding to transition captioning. |
| $\text{LM}_{\text{bwd}}$ | $I_{\text{post}},\ I_{\text{pre}},\ \underline{a}$ | Predicts the action text with the visual states reversed, testing whether gains arise from visual-state ordering rather than action-conditioned simulation. |

so that prediction units correspond to the lexical units under study. A language head tied to the text embedding matrix (Radford et al., 2019) predicts the next word, while a linear image head predicts the next patch embedding. Architectural dimensions, positional encodings, and component initialization are detailed in Appendix A.1.

### 2.2 Sequence Formats

Each training example contains a pre-condition image, an action text, and a post-condition image. At load time, we sample one of four sequence formats with fixed probabilities (Table 2). All formats share the same model parameters and the same causal next-element prediction objective. The supervision regime determines which positions contribute to the loss (Section 2.3). Serialization and sampling details in Appendix A.2.

### 2.3 Training Objectives

The total loss combines supervised next-token prediction for text positions and supervised next-embedding prediction for image positions:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{SIM}}. \tag{1}$$

Both terms are cross-entropy losses: $\mathcal{L}_{\text{LM}}$ is cross-entropy over tokens, while $\mathcal{L}_{\text{SIM}}$ is a visual simulation loss over patch embeddings.

**Language modeling loss.** For each supervised text position, the Transformer hidden state $h_t$ is passed through the language head to predict the next whole-word token:

$$p_\theta(x_{t+1} \mid x_{\le t}) = \text{softmax}(\text{lm_head}(h_t)). \tag{2}$$

The language modeling loss is standard causal cross-entropy over the supervised text positions:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{|\mathcal{S}_{\text{text}}|} \sum_{t \in \mathcal{S}_{\text{text}}} \log p_\theta(x_{t+1} \mid x_{\le t}). \tag{3}$$

The set $\mathcal{S}_{\text{text}}$ is determined by the sampled sequence format and supervision regime.

**Image prediction loss.** For each supervised image position, the model predicts the next patch embedding from the current hidden state:

$$\hat{z}_{t+1} = \text{img_head}(h_t), \tag{4}$$

The target embedding $z_{t+1}$ is produced by the same patch embedding module applied to the corresponding image patch. Our main image objective is an InfoNCE loss over supervised image positions. This keeps the language and image objectives on comparable scales without introducing additional loss weight. Given all supervised image predictions and targets in a batch, we $\ell_2$-normalize both sets of embeddings and compute the similarity matrix

$$s_{ij} = \hat{z}_i^\top z_j / \tau, \tag{5}$$

where $\tau$ is the temperature. The image loss is a cross-entropy loss that identifies the correct target embedding for each predicted embedding:

$$\mathcal{L}_{\text{SIM}} = -\frac{1}{|\mathcal{S}_{\text{SIM}}|} \sum_{i \in \mathcal{S}_{\text{SIM}}} \log \frac{\exp(s_{ii})}{\sum_{j \in \mathcal{S}_{\text{SIM}}} \exp(s_{ij})}. \tag{6}$$

Here, $\mathcal{S}_{\mathrm{SIM}}$ denotes the supervised image positions in the batch. This objective is the embedding-space analogue of next-token prediction: each predicted visual embedding must identify its corresponding next embedding from a batch-level set of candidates. It also avoids the mean-prediction behavior that can arise with regression losses on approximately zero-mean patch embeddings.

**Cosine alternative.** We also support a NEPA-style cosine-similarity objective:

$$\mathcal{L}_{\mathrm{SIM}}^{\cos} = -\frac{1}{|\mathcal{S}_{\mathrm{SIM}}|} \sum_{t \in \mathcal{S}_{\mathrm{SIM}}} \frac{\hat{z}_{t+1}^{\top} z_{t+1}}{\|\hat{z}_{t+1}\| \, \|z_{t+1}\|}. \tag{7}$$

This provides a direction-only objective, but it is not on the same cross-entropy scale as the language modeling loss. We therefore use InfoNCE as the main image objective unless otherwise specified.

### 2.4 Supervision Regimes

We consider two regimes for deciding which positions contribute to the loss.

**Full-sequence.** The default regime supervises all eligible text and image positions in the sampled sequence format. This is closest to standard pretraining on mixed visual-textual sequences: every supervised position contributes a next-token or next-embedding prediction loss, while the causal attention mask ensures that each prediction depends only on earlier context.

**Completion-only.** The completion-only regime supervises only the response block associated with the sampled format: the post-condition for forward simulation, the pre-condition for backward simulation, and the action text for language modeling. This is analogous to supervised fine-tuning, where the prompt is used as conditioning context but only the response contributes to the loss.

## 3 Experiment Setups

We use two corpora shown in Table 1. The real videos are skewed toward a long tail of fine-grained manipulation, while the world-model-distilled corpus complements it with broader coverage of locomotion, environment change, viewpoint shift, and appearance change. We therefore use the former as a real-data sanity check under a narrower distribution, and the latter for the broader generalization study. We present more details in Appendices B and C.

### 3.1 Real-Video Dataset

We first construct a real-video dataset from 5 open video-understanding datasets: Assembly101 (Sener et al., 2022), Ego4D (Grauman et al., 2022), EPIC-KITCHENS-100 (Damen et al., 2018; 2020), FineAction (Liu et al., 2022), and HoloAssist (Wang et al., 2023), which provide trimmed actions with labels and temporal boundaries. For each source, we keep the training split and repurpose the validation split as our evaluation set.

**Annotations.** We convert each annotated segment into a triple by taking the start frame as the pre-condition and the end frame as the post-condition. Since the original action labels are short verb–noun tags with varying granularity (e.g., *arrange (straighten, sort, distribute, align)* versus a bare *cut*), we convert each clean verb–noun label into a short vision-conditioned sentence given the pre-condition and post-condition frames. We use Gemini-3-Flash[1] for the best quality and throughput. This step standardizes the action text and serves as a consistency filter, as examples whose generated sentence fails to preserve the cleaned verb or noun are corrected or removed.

**Test splits.** Following Levin (2008), we assign verb lemmas to *manner* and *result* classes using a manually curated verb inventory. We then curate the test set by sampling 40 verbs from each class, with up to 50 examples per verb across available nouns. This subset allows us to test whether models rely more strongly on visual post-condition cues for result verbs.

### 3.2 World-Model-Distilled Dataset

While real videos provide real temporal observations, their action coverage is source-limited, and designing controlled experiments is very hard. To allow flexible ablations and broaden the range

[1] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

of action-conditioned transitions, we construct a complementary corpus from MSCOCO 2017 seed images (Lin et al., 2014) and a visual generative model as the world model.

**Annotations.** For each image, we use GPT-5.4-nano[2] to produce 4-5 diverse, physically and temporally plausible action candidates drawn from five action types: *object interaction*, *body motion*, *scene transition*, *state change*, *new element*. Each action candidate is converted into an editing instruction for FLUX.2 [klein]-9B (Labs, 2025), which generates the post-condition image and forms a triple with the source image and action text. We use GPT-5.5[3] to filter inconsistent triples, followed by manual quality verification. Generated triples are kept in the same split as their source images, with training-split triples used for training and test-split triples used for evaluation.

**Test splits.** We report results on two test splits. The $\mathrm{test\text{-}seen}$ split contains verb–object combinations observed during training. The $\mathrm{test\text{-}unseen}$ split contains combinations whose verb or object lemma is held out from training, and is used to evaluate generalization beyond memorized lexical pairs. Verb and object positions are annotated with a hybrid pipeline based on suffix matching, WordNet lemmatization, and a small hand-curated override table for irregular forms and plural-only nouns. Each test example contains a verb index, an object index, and canonical lemma identifiers, allowing surprisal to be aggregated by lemma.

### 3.3 Evaluation Protocol and Metrics

We evaluate word learning by measuring the *predictability* of action words and their participants from visual context. We use the forward language modeling format: the two visual states are provided before the action sentence, and the model is evaluated by next-token prediction at the annotated positions. When available, target words are mapped to canonical lemma identifiers, so that results can be aggregated by concept rather than only by surface form.

**Word surprisal.** Let $w_k$ denote a target word in the action sentence $\mathcal{S}$, such as an annotated verb or object token. Given the visual context $c = (\text{pre-condition}, \text{post-condition})$ and the preceding action prefix $a_{<k}$, we define target-word surprisal as

$$s_k(\mathcal{S}) = -\log_2 P(w_k \mid c, a_{<k}). \tag{8}$$

When multiple target words are annotated for the same example, we report their average surprisal:

$$s_{\text{target}}(\mathcal{S}) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} s_k(\mathcal{S}), \tag{9}$$

where $\mathcal{K}$ is the set of valid target positions.

**Learning curves and neural age of acquisition.** Following Ma et al. (2025), we summarize acquisition dynamics with both fitted learning curves and neural age of acquisition (nAoA). For each target word, we evaluate surprisal on the test set across training checkpoints and fit a learning curve over log training steps, capturing how predictability improves over the course of training. From the fitted curve, we compute nAoA as the training step at which the target crosses a specified acquisition threshold; lower nAoA indicates earlier acquisition. We report nAoA at the standard 0.50 threshold and, for robustness, also average nAoA over thresholds from 0.50 to 0.95.

## 4 Main Findings

Table 3 summarizes the main surprisal results at step $10{,}000$ on the world-model-distilled dataset, and Table 4 reports the corresponding evaluation on real trimmed action videos. Figure 2 and Table 5 complement the final-step results with learning-curve fits and nAoA results. Figure 3 shows qualitative examples. The central pattern is consistent across datasets: simulation is most useful as an auxiliary objective to language modeling, improving action-word and verb–noun predictability while leaving noun learning comparatively less affected.

**Simulation improves grounding when text supervision is sufficient.** The best-performing configurations allocate half the training budget to forward language modeling and half to simulation. In Table 3, these mixtures consistently improve verb and pair surprisal over the language-only baseline,

[2] https://developers.openai.com/api/docs/models/gpt-5.4-nano

[3] https://developers.openai.com/api/docs/models/gpt-5.5

Table 3: On the world-model-distilled dataset, surprisal results at training step 10,000, pooled over 3 random seeds. The four `Supervision` columns give the sampling probability of each base format: $\texttt{LM}_{\texttt{bwd}}$ (*postcon+precon→action*), $\texttt{LM}_{\texttt{fwd}}$ (*precon+postcon→action*), $\texttt{Sim}_{\texttt{bwd}}$ (*postcon+action→precon*), and $\texttt{Sim}_{\texttt{fwd}}$ (*precon+action→postcon*). The Generalization Gap is the aggregate pair-surprisal gap, Unseen−Seen.

| Supervision | | | | Loss Configuration | | Verb Surprisal ↓ | | Noun Surprisal ↓ | | Pair Surprisal ↓ | | Gen. Gap ↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\text{LM}_{\text{bwd}}$ | $\text{LM}_{\text{fwd}}$ | $\text{Sim}_{\text{bwd}}$ | $\text{Sim}_{\text{fwd}}$ | Applied To | Visual | Seen | Unseen | Seen | Unseen | Seen | Unseen | Pair (U−S) |
| 100% | – | – | – | Full Seq | InfoNCE | $9.06_{\pm.05}$ | $9.13_{\pm.04}$ | $2.50_{\pm.03}$ | $6.06_{\pm.04}$ | $5.78_{\pm.03}$ | $7.59_{\pm.03}$ | $+1.82_{\pm.04}$ |
| 50% | 50% | – | – | Full Seq | InfoNCE | $8.24_{\pm.06}$ | $8.60_{\pm.05}$ | $1.53_{\pm.03}$ | $4.68_{\pm.04}$ | $4.88_{\pm.03}$ | $6.64_{\pm.03}$ | $+1.76_{\pm.04}$ |
| – | 100% | – | – | Full Seq | InfoNCE | $8.31_{\pm.06}$ | $8.53_{\pm.05}$ | $1.55_{\pm.03}$ | $4.78_{\pm.04}$ | $4.93_{\pm.03}$ | $6.65_{\pm.03}$ | $+1.72_{\pm.04}$ |
| – | 50% | 50% | – | Full Seq | InfoNCE | $7.81_{\pm.05}$ | $8.15_{\pm.05}$ | $\mathbf{1.49}_{\pm.02}$ | $4.70_{\pm.04}$ | $4.65_{\pm.03}$ | $6.43_{\pm.03}$ | $+1.78_{\pm.04}$ |
| – | 50% | – | 50% | Full Seq | InfoNCE | $8.17_{\pm.05}$ | $8.41_{\pm.05}$ | $1.66_{\pm.03}$ | $5.01_{\pm.04}$ | $4.92_{\pm.03}$ | $6.71_{\pm.03}$ | $+1.79_{\pm.04}$ |
| – | 50% | 25% | 25% | Full Seq | InfoNCE | $\mathbf{7.66}_{\pm.05}$ | $\mathbf{8.05}_{\pm.05}$ | $1.56_{\pm.02}$ | $\mathbf{4.58}_{\pm.04}$ | $\mathbf{4.61}_{\pm.03}$ | $\mathbf{6.32}_{\pm.03}$ | $\mathbf{+1.71}_{\pm.04}$ |
| – | 33% | 33% | 33% | Full Seq | InfoNCE | $8.28_{\pm.06}$ | $8.61_{\pm.05}$ | $1.58_{\pm.03}$ | $4.83_{\pm.04}$ | $4.93_{\pm.03}$ | $6.72_{\pm.03}$ | $+1.79_{\pm.04}$ |
| 100% | – | – | – | Full Seq | Cosine | $9.60_{\pm.06}$ | $9.30_{\pm.05}$ | $2.42_{\pm.03}$ | $6.06_{\pm.04}$ | $6.01_{\pm.03}$ | $7.68_{\pm.03}$ | $\mathbf{+1.67}_{\pm.04}$ |
| 50% | 50% | – | – | Full Seq | Cosine | $7.85_{\pm.05}$ | $8.31_{\pm.04}$ | $1.62_{\pm.02}$ | $5.20_{\pm.04}$ | $4.73_{\pm.03}$ | $6.76_{\pm.03}$ | $+2.02_{\pm.04}$ |
| – | 100% | – | – | Full Seq | Cosine | $8.18_{\pm.05}$ | $8.59_{\pm.05}$ | $1.66_{\pm.03}$ | $5.30_{\pm.04}$ | $4.92_{\pm.03}$ | $6.94_{\pm.03}$ | $+2.02_{\pm.04}$ |
| – | 50% | 50% | – | Full Seq | Cosine | $\mathbf{7.30}_{\pm.05}$ | $\mathbf{7.89}_{\pm.04}$ | $\mathbf{1.58}_{\pm.02}$ | $5.16_{\pm.04}$ | $\mathbf{4.44}_{\pm.03}$ | $\mathbf{6.52}_{\pm.03}$ | $+2.08_{\pm.04}$ |
| – | 50% | – | 50% | Full Seq | Cosine | $8.23_{\pm.05}$ | $8.49_{\pm.04}$ | $1.76_{\pm.03}$ | $5.39_{\pm.04}$ | $5.00_{\pm.03}$ | $6.94_{\pm.03}$ | $+1.94_{\pm.04}$ |
| – | 50% | 25% | 25% | Full Seq | Cosine | $8.32_{\pm.05}$ | $8.57_{\pm.05}$ | $1.67_{\pm.03}$ | $\mathbf{5.11}_{\pm.04}$ | $5.00_{\pm.03}$ | $6.84_{\pm.03}$ | $+1.85_{\pm.04}$ |
| – | 33% | 33% | 33% | Full Seq | Cosine | $7.68_{\pm.05}$ | $8.16_{\pm.04}$ | $1.69_{\pm.03}$ | $5.18_{\pm.04}$ | $4.68_{\pm.03}$ | $6.67_{\pm.03}$ | $+1.98_{\pm.04}$ |
| 100% | – | – | – | Completion | – | $9.58_{\pm.06}$ | $9.39_{\pm.05}$ | $2.35_{\pm.03}$ | $6.12_{\pm.04}$ | $5.97_{\pm.03}$ | $7.76_{\pm.03}$ | $\mathbf{+1.79}_{\pm.04}$ |
| 50% | 50% | – | – | Completion | – | $9.13_{\pm.06}$ | $9.06_{\pm.05}$ | $1.78_{\pm.03}$ | $5.65_{\pm.04}$ | $5.46_{\pm.03}$ | $7.36_{\pm.03}$ | $+1.90_{\pm.04}$ |
| – | 100% | – | – | Completion | – | $8.25_{\pm.05}$ | $8.58_{\pm.05}$ | $1.68_{\pm.02}$ | $5.32_{\pm.04}$ | $4.97_{\pm.03}$ | $6.95_{\pm.03}$ | $+1.98_{\pm.04}$ |
| – | 50% | 50% | – | Completion | InfoNCE | $7.76_{\pm.05}$ | $\mathbf{8.31}_{\pm.05}$ | $1.69_{\pm.02}$ | $4.96_{\pm.04}$ | $4.72_{\pm.03}$ | $6.64_{\pm.03}$ | $+1.91_{\pm.04}$ |
| – | 50% | – | 50% | Completion | InfoNCE | $\mathbf{7.75}_{\pm.05}$ | $8.36_{\pm.04}$ | $1.69_{\pm.02}$ | $4.94_{\pm.04}$ | $\mathbf{4.72}_{\pm.03}$ | $6.65_{\pm.03}$ | $+1.93_{\pm.04}$ |
| – | 50% | 25% | 25% | Completion | InfoNCE | $7.82_{\pm.05}$ | $8.33_{\pm.05}$ | $\mathbf{1.65}_{\pm.02}$ | $\mathbf{4.82}_{\pm.03}$ | $4.73_{\pm.03}$ | $\mathbf{6.58}_{\pm.03}$ | $+1.84_{\pm.04}$ |
| – | 33% | 33% | 33% | Completion | InfoNCE | $8.59_{\pm.05}$ | $8.94_{\pm.05}$ | $1.87_{\pm.03}$ | $5.39_{\pm.04}$ | $5.23_{\pm.03}$ | $7.17_{\pm.03}$ | $+1.94_{\pm.04}$ |

Table 4: On the real-video test set, Full-sequence + InfoNCE loss, pooled over 3 random seeds.

| Supervision | | | Manner Surprisal ↓ | | Result Surprisal ↓ | |
|---|---|---|---|---|---|---|
| $\text{LM}_{\text{fwd}}$ | $\text{Sim}_{\text{bwd}}$ | $\text{Sim}_{\text{fwd}}$ | Verb | Noun | Verb | Noun |
| 100% | – | – | $3.30_{\pm.02}$ | $2.45_{\pm.02}$ | $3.94_{\pm.02}$ | $3.22_{\pm.04}$ |
| 50% | 50% | – | $3.21_{\pm.00}$ | $2.34_{\pm.02}$ | $3.81_{\pm.00}$ | $3.05_{\pm.01}$ |
| 50% | – | 50% | $3.27_{\pm.04}$ | $2.40_{\pm.01}$ | $3.81_{\pm.01}$ | $3.08_{\pm.01}$ |
| 50% | 25% | 25% | $\mathbf{3.15}_{\pm.02}$ | $\mathbf{2.28}_{\pm.02}$ | $\mathbf{3.66}_{\pm.01}$ | $\mathbf{2.97}_{\pm.01}$ |

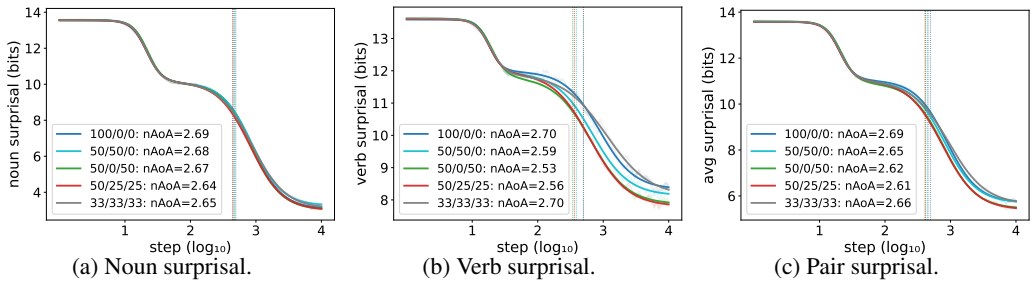
Table 5: nAoA values for Full-sequence + InfoNCE from a single double-sigmoid fit per (mix, metric, split) to the 3-seed-pooled trajectory (step $\leq$ 10,000). `@[.5]`: $\log_{10}$(step) at which the fitted curve crosses the midpoint of the observed range. `@[.5:.95]`: average of the same crossing across progress thresholds $\{0.50, 0.55, \ldots, 0.95\}$. $\Delta$ denotes improvement over the 100% LM baseline (current setting − baseline), so more negative values indicate larger gains. `All` pools seen+unseen combos into one trajectory before fitting.

| Supervision | | | Verb nAoA ↓ | | | | | | Noun nAoA ↓ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Seen | | Unseen | | All | | Seen | | Unseen | | All | |
| $\text{LM}_{\text{fwd}}$ | $\text{Sim}_{\text{bwd}}$ | $\text{Sim}_{\text{fwd}}$ | @[.5] | @[.5:.95] | @[.5] | @[.5:.95] | @[.5] | @[.5:.95] | @[.5] | @[.5:.95] | @[.5] | @[.5:.95] | @[.5] | @[.5:.95] |
| 100% | – | – | 2.73 | 3.13 | 2.66 | 3.11 | 2.70 | 3.12 | 2.61 | 2.96 | 2.81 | 3.17 | 2.69 | 3.05 |
| 50% | 50% | – | **2.54** | **2.98** | **2.53** | **2.98** | **2.53** | **2.97** | 2.61 | 2.96 | **2.75** | **3.13** | 2.67 | 3.03 |
| 50% | – | 50% | 2.62 | 3.04 | 2.57 | 3.01 | 2.59 | 3.03 | 2.62 | 2.97 | 2.77 | 3.16 | 2.68 | 3.04 |
| 50% | 25% | 25% | 2.58 | 3.00 | 2.54 | **2.98** | 2.56 | 2.99 | **2.58** | **2.94** | **2.75** | 3.14 | **2.64** | **3.02** |
| 33% | 33% | 33% | 2.77 | 3.23 | 2.63 | 3.15 | 2.70 | 3.19 | **2.58** | 2.96 | 2.77 | 3.17 | 2.65 | 3.05 |
| Best $\Delta$ vs. LM | | | −0.19 | −0.15 | −0.13 | −0.13 | −0.17 | −0.15 | −0.03 | −0.02 | −0.06 | −0.04 | −0.05 | −0.03 |

with the balanced bidirectional mix performing best. By contrast, the symmetric mixture with only one third language modeling is weaker, both in final surprisal and in acquisition speed. This suggests that simulation supplies useful visual structure, but the model still needs substantial direct exposure to action prediction to convert that structure into lexical gains.

**Backward simulation contributes more than forward simulation.** Although balanced bidirectional simulation gives the strongest final surprisal in Table 3, backward simulation gives the clearest improvement in verb acquisition speed in Table 5. This suggests that inverse prediction is a more diagnostic auxiliary task for action meaning (Baker et al., 2009): recovering the pre-condition from the post-condition and action requires representing what the action changed, whereas forward prediction can partly exploit visual continuity between the two states.

**The pattern transfers to real action videos.** The real-video results in Table 4 show that the same principle extends beyond the controlled before/after setting. Despite the noisier visual input, adding




(a) Noun surprisal. (b) Verb surprisal. (c) Pair surprisal.

Figure 2: Fitted learning curves for the Full-sequence + InfoNCE loss configuration, comparing five supervision mixes (legend: `Fwd-LM/Fwd-Sim/Bwd-Sim` sampling %). Each curve is a double-sigmoid fit to the mean surprisal trajectory pooled over 3 random seeds and seen+unseen test combos. Vertical dotted lines mark `nAoA@0.50`, the $\log_{10}(\text{step})$ at which each curve crosses the midpoint of its observed range.

Table 6: Manner vs result verb classes, Full-sequence + InfoNCE loss (step $\leq$ 10,000). **Final verb surprisal**: per-verb mean (pooled over combos and 3 seeds) at the final step, then averaged across verbs in the class. **nAoA@[.5:.95]**: average crossing across progress thresholds $\{0.50, 0.55, \ldots, 0.95\}$, averaged across verbs that fit successfully. Learning curves are fit using approximately 57K examples from the test set. `Supervision` columns give the sampling probability of each base format. $n_{\text{manner}} = 34$, $n_{\text{result}} = 54$ for surprisal; $n_{\text{manner}} = 32$, $n_{\text{result}} = 50$ for nAoA. $\Delta$ denotes the difference from the 100% forward-LM baseline (current setting $-$ baseline), so more negative values indicate larger gains.

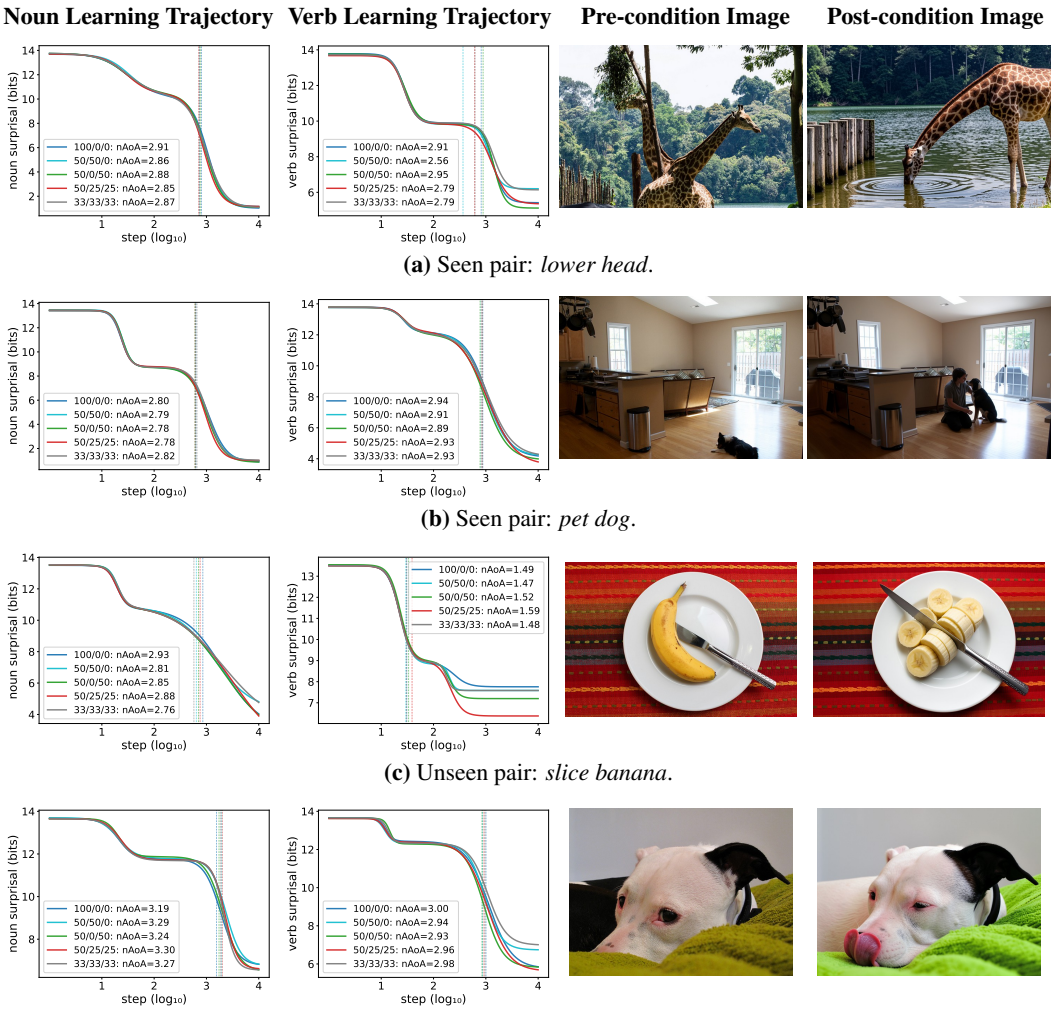
| Supervision | | | Surprisal ↓ | | | | nAoA@[.5:.95] ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\text{LM}_{\text{fwd}}$ | $\text{Sim}_{\text{bwd}}$ | $\text{Sim}_{\text{fwd}}$ | Man. | $\Delta$ | Res. | $\Delta$ | Man. | $\Delta$ | Res. | $\Delta$ |
| 100% | – | – | 8.14 | – | 8.84 | – | 2.99 | – | 2.89 | – |
| 50% | 50% | – | 7.68 | −0.46 | 8.41 | −0.43 | 3.00 | +0.01 | **2.74** | **-0.15** |
| 50% | – | 50% | 7.99 | −0.15 | 8.72 | −0.12 | 3.01 | +0.02 | 2.80 | −0.09 |
| 50% | 25% | 25% | **7.56** | **-0.58** | **8.21** | **-0.63** | **2.97** | **-0.02** | 2.75 | −0.14 |
| 33% | 33% | 33% | 8.29 | +0.15 | 8.79 | −0.05 | 2.99 | 0.00 | 2.95 | +0.06 |

simulation improves final lexical predictability, and the balanced bidirectional mixture performs best across the reported manner/result verb and noun columns. This strengthens the main claim: visual simulation is not merely exploiting artifacts of a controlled transition dataset, but provides a useful auxiliary signal for grounding action language in naturalistic visual observations.

**Simulation accelerates verb learning, but not noun learning.** Table 5 shows that simulation advances verb nAoA by 0.13–0.19 $\log_{10}$ steps, while noun nAoA changes by only 0.02–0.06. Figure 2 shows the same pattern visually. Nouns can often be grounded by direct reference to visible entities, whereas verbs depend on the relation between pre-condition and post-condition.

**Result verbs benefit more from simulation than manner verbs.** Table 6 shows the clearest difference in acquisition speed. Across the three 50%-LM simulation settings, result-verb nAoA has $\Delta = -0.15, -0.09, -0.14$, compared with $\Delta = +0.01, +0.02, -0.02$ for manner verbs. Final surprisal improves for both classes, with balanced bidirectional simulation also giving a slightly larger gain for result verbs ($-0.63$ vs. $-0.58$). This is consistent with result verbs being more directly tied to visible state changes, while manner verbs depend more on how an action unfolds over time. This pattern is consistent with the semantic distinction between the two classes. Result verbs are closely tied to visible changes of state, which are directly exposed by before/after prediction. Manner verbs depend more on how an action unfolds over time, so a static transition signal is useful but incomplete.

**Alternative losses change the winner, not the conclusion.** The remaining blocks of Table 3 show that the qualitative result is stable across loss and supervision variants. The exact best row changes with the visual loss, supervision format, and metric, but simulation generally improves over language-only training on verb and pair surprisal. We therefore do not interpret any single supervision mix as universally optimal. The stable conclusion is narrower and stronger: action-conditioned visual simulation improves lexical grounding when used as an auxiliary objective, especially for predicates and compositional verb–noun prediction.




**(a)** Seen pair: *lower head*.

**(b)** Seen pair: *pet dog*.

**(c)** Unseen pair: *slice banana*.

**(d)** Unseen pair: *lick nose*.

Figure 3: Examples for two seen and two unseen verb-noun pairs. Rows correspond to seen result verb *lower head*, seen manner verb *pet dog*, unseen result verb *slice banana*, and unseen manner verb *lick nose*. Columns show the noun/verb surprisal learning curves and the pre-condition/post-condition images. Curves pool 3 random seeds under the full-sequence setting with InfoNCE loss.

## 5 Related Work

### 5.1 Causal perception and world modeling

A long tradition in cognitive science treats human reasoning as grounded in internal models of the world: structured representations that support prediction, counterfactual projection, causal explanation, and action planning (Johnson-Laird, 1983; Nersessian, 1993; Wooldridge & Jennings, 1995; Kappes & Morewedge, 2016; Russell & Norvig, 2016). Developmental evidence suggests that this event structure emerges early: infants distinguish causal from non-causal events, represent event roles before they have a mature lexicon (Leslie & Keeble, 1987; Spelke, 1990; Mandler, 1992; Oakes & Cohen, 1990; Cohen et al., 1998). Since these roles are also central to verb meaning (Cohen et al., 1999; Solstad & Bott, 2017), causal perception therefore links visual event understanding to action-word learning.

Computer vision has studied related questions through object affordances, human-object interactions, and causal relations between actions and object states (Koppula et al., 2013; Fathi & Rehg, 2013; Yang et al., 2013; Wang et al., 2016; Zhuo et al., 2019). Other work infers object states from observations (Fire & Zhu, 2015; 2017; Alayrac et al., 2017; Liu et al., 2017; Hong et al., 2021), predicts future state changes (Zhou & Berg, 2016; Xiong et al., 2018; Gao et al., 2018), or jointly

models physical state and dynamics (Wu et al., 2015; 2016; 2017). More recent world-modeling approaches further emphasize prediction as a central objective for agents.

Despite this progress, most prior work treats state change primarily as a visual or physical modeling problem, leaving the connection to lexical acquisition less explored.

### 5.2 Grounded word acquisition

Grounded word learning has traditionally been framed as the problem of mapping words to referents in the world (Harnad, 1990; Gleitman & Landau, 1994; Clark, 1995). Early computational and visual grounding work modeled word-to-symbol mappings, word-object alignment, and grounding in richer visual features and structured scene representations (Siskind, 1996; Regier, 2005; Goodman et al., 2007; Fazly et al., 2010; Roy & Pentland, 2002; Yu, 2005; Xu & Tenenbaum, 2007; Yu & Ballard, 2007; Yu & Siskind, 2013; Qu & Chai, 2010; Mao et al., 2019; 2021; Pratt et al., 2020). Large-scale vision-language models have extended this paradigm to region-level (Li et al., 2022; Ma et al., 2023; Chen et al., 2023; You et al., 2024; Wang et al., 2024) and pixel-level (Xia et al., 2024; Rasheed et al., 2024; Zhang et al., 2024) grounding, with strong performance on referring expression comprehension (Chen et al., 2024).

Despite this progress, much of this literature remains object-centered. This formulation is well suited to nouns but incomplete for verbs, which often denote event structure over time. Developmental work indicates that objects are stable perceptual units, whereas verbs require the learner to identify event components (Gentner, 1982; 2006; Golinkoff et al., 1987). Verb learning is also shaped by the nouns that participate in the event (Arunachalam & Waxman, 2011; Ferguson et al., 2014). Together, these findings suggest that lexical acquisition couples object identity with action structure.

Our work studies this coupling in neural models. Rather than treating grounding as static word-referent alignment, we ask whether lexical learning improves under action-conditioned visual transitions. This framing suggests that transition prediction is useful for verbs, especially result verbs.

### 5.3 Psychologically motivated analysis of language models

Neural language models differ from human learners in data, objectives, architectures, and inductive biases (Baroni, 2022; Warstadt & Bowman, 2022). Nevertheless, they can serve as controlled computational systems for testing which learning signals support particular linguistic behaviors. Prior work has connected language-model surprisal to psycholinguistic variables and lexical acquisition patterns (Portelance et al., 2020; Ma et al., 2023), and has examined how linguistic knowledge emerges over training time (Sellam et al., 2021; Blevins et al., 2022; Biderman et al., 2023; Xia et al., 2023; Chang & Bergen, 2022; Chang et al., 2023; Evanson et al., 2023; Ma et al., 2025).

If models are used as such scientific instruments, benchmark outcomes alone are insufficient (Baroni, 2022; Portelance, 2022). The relevant question is which training factors make success possible (Warstadt & Bowman, 2022; Portelance, 2022; Portelance & Jasbi, 2023). We therefore vary visual simulation supervision, measure lexical surprisal over training, and compare acquisition dynamics across nouns, verbs, manner verbs, and result verbs.

## 6 Discussion and Limitations

Our framework models single-step action-effect transitions and therefore does not yet test longer-horizon compositional reasoning over sequences of state changes. The controlled transition corpus is also generated using language and image generative models, so its action distribution inevitably reflects the generator's priors over plausible actions, participants, and resulting states. On the language side, we use a fixed-vocabulary whitespace tokenizer so that target verbs and nouns map to single tokens, which simplifies lexical surprisal evaluation but precludes subword generalization to related or unseen forms. Finally, our experiments use English captions and action descriptions only, and therefore do not address cross-linguistic variation in how events and state changes are lexicalized.

## AI USE STATEMENT

We used generative AI tools for dataset construction and annotation, including the generation and filtering of synthetic action–state transition data, and for assistance with software implementation and code development. All AI-assisted code was reviewed and tested by the authors, and the resulting datasets were subject to author-designed filtering and quality-control procedures.

We did not use generative AI to develop the core conceptual framework, formulate or refine the research hypotheses, design the research methodology or experiments, formulate mathematical claims or proofs, or interpret the experimental results; these decisions and analyses were made by the authors.

Additionally, we used generative AI for final-stage language and grammar editing to improve the readability of the manuscript. All AI-assisted edits were reviewed by the authors. We take responsibility for the final content of this work, including its data, code, analyses, claims, and text.

## REFERENCES

Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2127–2136, 2017.

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

Sudha Arunachalam and Sandra R Waxman. Grammatical form and semantic context in verb learning. *Language Learning and Development*, 7(3):169–184, 2011.

Chris L Baker, Rebecca Saxe, and Joshua B Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning, 2019. URL `https://arxiv.org/abs/1908.05656`.

Marco Baroni. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language*, pp. 1–16, 2022.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3575–3590, 2022.

Kyle Buettner and Adriana Kovashka. Investigating the role of attribute context in vision-language models for object recognition and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5474–5484, 2024.

Tyler A Chang and Benjamin K Bergen. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16, 2022.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *arXiv preprint arXiv:2308.15419*, 2023.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Eve V Clark. *The lexicon in acquisition*. Number 65. Cambridge University Press, 1995.

L. Cohen, G. Amsel, M. A. Redford, and M. Casasola. The development of infant causal perception. In A. Slater (ed.), *Perceptual development: Visual, auditory and speech perception in infancy*. Psychology Press, 1999.

Leslie B Cohen, Geoffrey Amsel, Melissa A Redford, and Marianella Casasola. The development of infant causal perception. *Perceptual development: Visual, auditory, and speech perception in infancy*, pp. 167–209, 1998.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision. *CoRR*, abs/2006.13256, 2020.

Linnea Evanson, Yair Lakretz, and Jean-Rémi King. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12205–12218, 2023.

Alireza Fathi and James M Rehg. Modeling actions through state changes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2579–2586, 2013.

Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063, 2010.

Brock Ferguson, Eileen Graf, and Sandra R Waxman. Infants use known verbs to learn novel nouns: Evidence from 15-and 19-month-olds. *Cognition*, 131(1):139–146, 2014.

Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23, 2015.

Amy. Fire and Song-Chun. Zhu. Inferring hidden statuses and actions in video by causal reasoning. In *CVPR Workshop: Vision Meets Cognition*, 2017.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024.

Qiaozi Gao, Malcolm Doering, Shaohua Yang, and Joyce Chai. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016. doi: 10.18653/v1/p16-1171. URL https://doi.org/10.18653/v1/p16-1171.

Qiaozi Gao, Shaohua Yang, Joyce Chai, and Lucy Vanderwende. What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 934–945, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, Xinqi Huang, Xinyu Fang, Lu Sun, Gautham Kishore, Bo Ai, Stone Tao, Mengyang Liu, Jiaxi Yang, Chao-Jung Lai, Chuanyang Jin, Jiannan Xiang, Benhao Huang, Zeming Chen, David Danks, Hao Su, Tianmin Shu, Ziqiao Ma, Lianhui Qin, and Zhiting Hu. Do vision-language models have internal world models? towards an atomic evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Dedre Gentner. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *Center for the Study of Reading Technical Report; no. 257*, 1982.

Dedre Gentner. Why verbs are hard to learn. *Action meets word: How children learn verbs*, pp. 544–564, 2006.

Lila R Gleitman and Barbara Landau. *The acquisition of the lexicon*. MIT Press, 1994.

Roberta Michnick Golinkoff, Kathryn Hirsh-Pasek, Kathleen M Cauley, and Laura Gordon. The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, 14(1): 23–45, 1987.

Noah Goodman, Joshua Tenenbaum, and Michael Black. A bayesian framework for cross-situational word-learning. *Advances in neural information processing systems*, 20, 2007.

Google. Gemini: A family of highly capable multimodal models. Technical report, Google, 2023.

Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10514–10525. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/76dc611d6ebaafc66cc0879c71b5db5c-Paper.pdf.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

Xin Hong, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Transformation driven visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6903–6912, June 2021.

P.N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press, 1983. ISBN 9780674568822. URL https://books.google.com/books?id=FS3zSKAfLGMC.

Heather Barry Kappes and Carey K Morewedge. Mental simulation as substitute for experience. *Social and Personality Psychology Compass*, 10(7):405–420, 2016.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics.

Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Alan M Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3): 265–288, 1987.

Beth Levin. A constraint on verb meanings: Manner/result complementarity. *Cognitive Science Department Colloqium Series, Brown University, Providence, RI, March*, 17:2008, 2008.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

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*, pp. 740–755. Springer, 2014.

Yang Liu, Ping Wei, and Song-Chun Zhu. Jointly recognizing object fluents and tasks in egocentric videos. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2943–2951, 2017.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31:6937–6950, 2022. doi: 10.1109/TIP.2022.3217368.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 524–544, 2023.

Ziqiao Ma, Zekun Wang, and Joyce Chai. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 991–1010, 2025.

Jean M Mandler. How to build a baby: Ii. conceptual primitives. *Psychological review*, 99(4):587, 1992.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, sentences from natural supervision. *International Conference on Learning Representations (ICLR)*, 2019.

Jiayuan Mao, Freda H. Shi, Jiajun Wu, Roger P. Levy, and Joshua B. Tenenbaum. Grammar-based grounded lexicon learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Nancy J Nersessian. In the theoretician's laboratory: Thought experimenting as mental modeling. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume 1992, pp. 291–301. Cambridge University Press, 1993.

Lisa M Oakes and Leslie B Cohen. Infant perception of a causal event. *Cognitive Development*, 5(2): 193–207, 1990.

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

Eva Portelance. *Neural Network Approaches to the Study of Word Learning*. PhD thesis, Stanford University, 2022.

Eva Portelance and Masoud Jasbi. The roles of neural networks in language acquisition. *PsyArXiv*, 2023.

Eva Portelance, Judith Degen, and Michael C Frank. Predicting age of acquisition in early word learning using recurrent neural networks. In *CogSci*, 2020.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pp. 314–332. Springer, 2020.

Shaolin Qu and Joyce Yue Chai. Context-based word acquisition for situated dialogue in a virtual world. *Journal of Artificial Intelligence Research*, 37:247–277, 2010.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Terry Regier. The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6):819–865, 2005.

Deb K Roy and Alex P Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, et al. The multiberts: Bert reproductions for robustness analysis. In *International Conference on Learning Representations*, 2021.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2022.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models, 2024. URL https://arxiv.org/abs/2410.23266.

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

Torgrim Solstad and Oliver Bott. Causality and causal reasoning in natural language. *The Oxford Handbook of Causal Reasoning*, pp. 619–644, 2017.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2658–2667, 2016.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20270–20281, October 2023.

Alex Warstadt and Samuel R Bowman. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pp. 17–60, 2022.

Michael Wooldridge and Nicholas R Jennings. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152, 1995.

Jiajun Wu, Ilker Yildirim, Joseph J Lim, William T Freeman, and Joshua B Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*, pp. 127–135, 2015.

Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, volume 2, pp. 7, 2016.

Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual de-animation. In *Advances in Neural Information Processing Systems*, pp. 152–163, 2017.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13711–13738, 2023.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Fei Xu and Joshua B Tenenbaum. Word learning as bayesian inference. *Psychological review*, 114 (2):245, 2007.

Sihan Xu, Ziqiao Ma, Wenhao Chai, Xuweiyi Chen, Weiyang Jin, Joyce Chai, Saining Xie, and Stella X Yu. Next-embedding prediction makes strong vision learners. *arXiv preprint arXiv:2512.16922*, 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis L Brown II, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2025.

Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 2563–2570. IEEE, 2013.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

Chen Yu. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection science*, 17(3-4):381–397, 2005.

Chen Yu and Dana H Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.

Haonan Yu and Jeffrey Mark Siskind. Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 53–63, 2013.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pp. 69–85. Springer, 2016.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yipin Zhou and Tamara L Berg. Learning temporal transformations from time-lapse videos. In *European Conference on Computer Vision*, pp. 262–277. Springer, 2016.

Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning via prior knowledge and state transitions. *ACM Multimedia*, 2019.

# A METHOD

## A.1 ARCHITECTURE

We initialize the visual backbone from NEPA (Xu et al., 2025), an autoregressive visual model pretrained with causal patch-embedding prediction, while initializing the language-specific components from scratch. This gives the model visual predictive structure at initialization, but requires it to acquire the mapping between words and visually grounded state changes during training.

**Tokenization.** We use whole-word rather than subword tokenization because the object of study is word learning, and the relevant lexical units are complete words rather than subword fragments.

**Backbone.** We use the `NEPA-large` backbone, which has 24 Transformer layers, hidden size 1024, and 16 attention heads, pretrained on ImageNet-1K. Image inputs are $224 \times 224$ tensors and are mapped to $N = 256$ patch embeddings by a convolutional patch embedding layer. Text inputs are represented with whole-word tokenization over the corpus. The original NEPA model uses two-dimensional rotary positional embeddings tied to image-grid coordinates. Because our sequences interleave image patches, text tokens, and boundary tokens, we replace these with learned one-dimensional absolute positional embeddings shared across all sequence positions. We also add token-type embeddings to distinguish image patches, text tokens, and boundary tokens. The Transformer trunk is shared across modalities and sequence formats.

**Output heads.** The model has two output heads. The language head predicts the next whole-word token and is tied to the text embedding matrix (Radford et al., 2019). The image head is a linear projection that predicts the next visual patch embedding. We initialize the patch embedding layer, Transformer trunk, and final layer normalization from `NEPA-large`. The text embeddings, token-type embeddings, absolute positional embeddings, and language head are randomly initialized.

## A.2 SEQUENCE FORMATS

Each training example contains a pre-condition image, an action text, and a post-condition image. At load time, we randomly sample one sequence format for each example according to fixed probabilities. All formats share the same model parameters and the same causal next-element prediction objective. They differ only in the order of the visual and textual components and in which positions are included in the loss. Special tokens `<pre>`, `</pre>`, `<post>`, and `</post>` mark the beginning and end of the pre-condition and post-condition image blocks in the actual input sequence.

**Forward simulation.** In the forward simulation format, the pre-condition image and action text appear before the post-condition image. This corresponds to **pre-condition** + **action** $\rightarrow$ **post-condition**: the model predicts the resulting visual state from an initial state and an action description.

**Backward simulation.** In the backward simulation format, the post-condition image and action text appear before the pre-condition image. This corresponds to **post-condition** + **action** $\rightarrow$ **pre-condition**: the model infers a compatible prior visual state from an observed result and an action description. Like forward simulation, this objective grounds action words in visually represented state changes; empirically, it can provide a stronger learning signal than forward simulation.

**Forward language modeling.** In the forward language modeling format, the pre-condition and post-condition images appear before the action text. This corresponds to **pre-condition** + **post-condition** $\rightarrow$ **action**: the model predicts the action text associated with an observed visual transition, i.e., video captioning.

Table 7: Parameters in the available generation code and run templates.

| Parameter | Available configuration |
|---|---|
| Model | FLUX.2 [klein]-9B |
| Denoising steps | 4 |
| Guidance scale | 1.0 |
| Edited images per action | 1 |

**Backward language modeling.** In the backward language modeling format, the post-condition and pre-condition images appear before the action text. This corresponds to **post-condition + pre-condition → action**. We use this to reject the alternative hypothesis that improvements come from the ordering of visual states in language modeling, rather than from visual simulation objectives that require image-embedding prediction.

**Format sampling.** The sampling probabilities over sequence formats determine how the training budget is allocated between language modeling and visual simulation. A pure language-modeling baseline assigns all examples to language-modeling formats, while simulation-augmented settings reserve part of the budget for image-embedding prediction in one or both directions. Our default setting assigns 50% of examples to language modeling and 50% to simulation objectives. In Section 4, we vary this allocation to measure how different mixtures of forward language modeling, backward language modeling, forward simulation, and backward simulation affect word learning.

# B Dataset

## B.1 World-model-distilled dataset

**Generation procedure.** Our generation pipeline begins with an MSCOCO 2017 source image and requests four to five candidate future actions from GPT-5.4-nano. For each candidate, the structured response contains an action phrase, its moving subjects, an expected visual change, an action type, an optional viewpoint change, and a corresponding image-editing instruction. The five action types are *object interaction*, *body motion*, *scene transition*, *state change*, and *new element*. The prompt requests at least three action types, at least one new-element candidate, and at least two event-linked viewpoint changes per source. It emphasizes physically and temporally plausible, visibly distinguishable outcomes and preservation of relevant object identities, materials, support relations, and scene content. These are prompting constraints, rather than independently measured properties of every generated sample.

Each candidate's editing instruction is applied independently to the same source image with FLUX.2 [klein]-9B. The action phrase becomes the training text, while the expected-change field is retained as auxiliary metadata. This field is predicted before image editing and is not a separate caption of the realized output image.

**Image-generation settings.** Table 7 reports the generation configuration. We preserve the source canvas by padding the right and bottom edges to multiples of 16 using edge pixels, generate the edited image, and crop back to the original dimensions. The implementation accepts inputs with both dimensions at least 64 pixels and a total area at most four megapixels.

**Structured validation and curation.** The pipeline validates the generated JSON structure and requires a nonempty action phrase, expected-change description, and exactly one editing prompt for each action index. Invalid batch responses are logged and rerun.

**Splits and lexical metadata.** Table 8 describes the statistics of the world-model-distilled dataset. The two test subsets used for the main experiments are curated from the full test pool.

**Vocabulary counting.** Vocabulary sizes count distinct non-missing annotation labels within each split. Verb counts use the verb labels. Noun vocabularies combine the subject and object labels.

Table 8: World-model-distilled statistics. Seen and unseen are curated subsets of the full test pool. Nouns count distinct stored subject and object labels jointly, excluding missing values.

| Split | Triples | Verbs | Nouns |
|---|---|---|---|
| Train | 488,703 | 1,788 | 5,322 |
| Validation | 20,696 | 785 | 1,924 |
| Full test pool | 168,690 | 1,338 | 3,863 |
| Test (seen) | 47,332 | 97 | 1,220 |
| Test (unseen) | 12,097 | 120 | 794 |

Table 9: Real-video statistics by source and split. Train and Test (curated) use the same split definitions as the main-text table. Test (curated) is a 3,999-example subset of the full evaluation pool. Vocabulary sizes count distinct nonmissing verb and object labels within each row; All sources takes the union across sources within that split. The full evaluation pool includes 95,690 source-validation examples and 21,893 Assembly101 official-test examples.

| Source | Split | Triples | Verbs | Nouns |
|---|---|---|---|---|
| Assembly101 | Train | 47,535 | 16 | 58 |
| | Test (curated) | 393 | 11 | 38 |
| | Full eval. pool | 37,565 | 16 | 60 |
| Ego4D FHO-LTA | Train | 63,846 | 116 | 460 |
| | Test (curated) | 1,598 | 46 | 247 |
| | Full eval. pool | 33,001 | 108 | 419 |
| EPIC-KITCHENS-100 | Train | 67,109 | 342 | 699 |
| | Test (curated) | 457 | 32 | 129 |
| | Full eval. pool | 9,646 | 168 | 359 |
| FineAction | Train | 47,852 | 45 | 64 |
| | Test (curated) | 990 | 32 | 49 |
| | Full eval. pool | 19,474 | 44 | 63 |
| HoloAssist | Train | 124,702 | 45 | 102 |
| | Test (curated) | 561 | 22 | 59 |
| | Full eval. pool | 17,897 | 46 | 98 |
| All sources | Train | 351,044 | 400 | 1,016 |
| | Test (curated) | 3,999 | 80 | 407 |
| | Full eval. pool | 117,583 | 257 | 724 |

Annotation strings are counted without additional normalization. Vocabulary sizes pooled across sources are computed from the union of their label sets.

**Illustrative generated pairs.** Figure 4 shows eight generated pairs sampled from the world-model-distilled dataset.

### B.2 Real-video dataset

**Sources and temporal boundaries.** The corpus pools Assembly101, Ego4D FHO-LTA, EPIC-KITCHENS-100, FineAction, and HoloAssist. An example uses the annotated segment's start and end boundaries as the pre-condition and post-condition observations. The extraction retains each source's timestamp or frame-index convention.

**Action annotation and splits.** Each action is described by an image-conditioned Gemini sentence and paired with cleaned verb and object labels. Lexical alignment provides token positions and matched surface forms. Table 9 gives the statistics for each source and split.

**Manner and result inventory.** The operational classification uses a manually curated lemma inventory. Manner verbs describe how an action is performed, such as its characteristic motion or

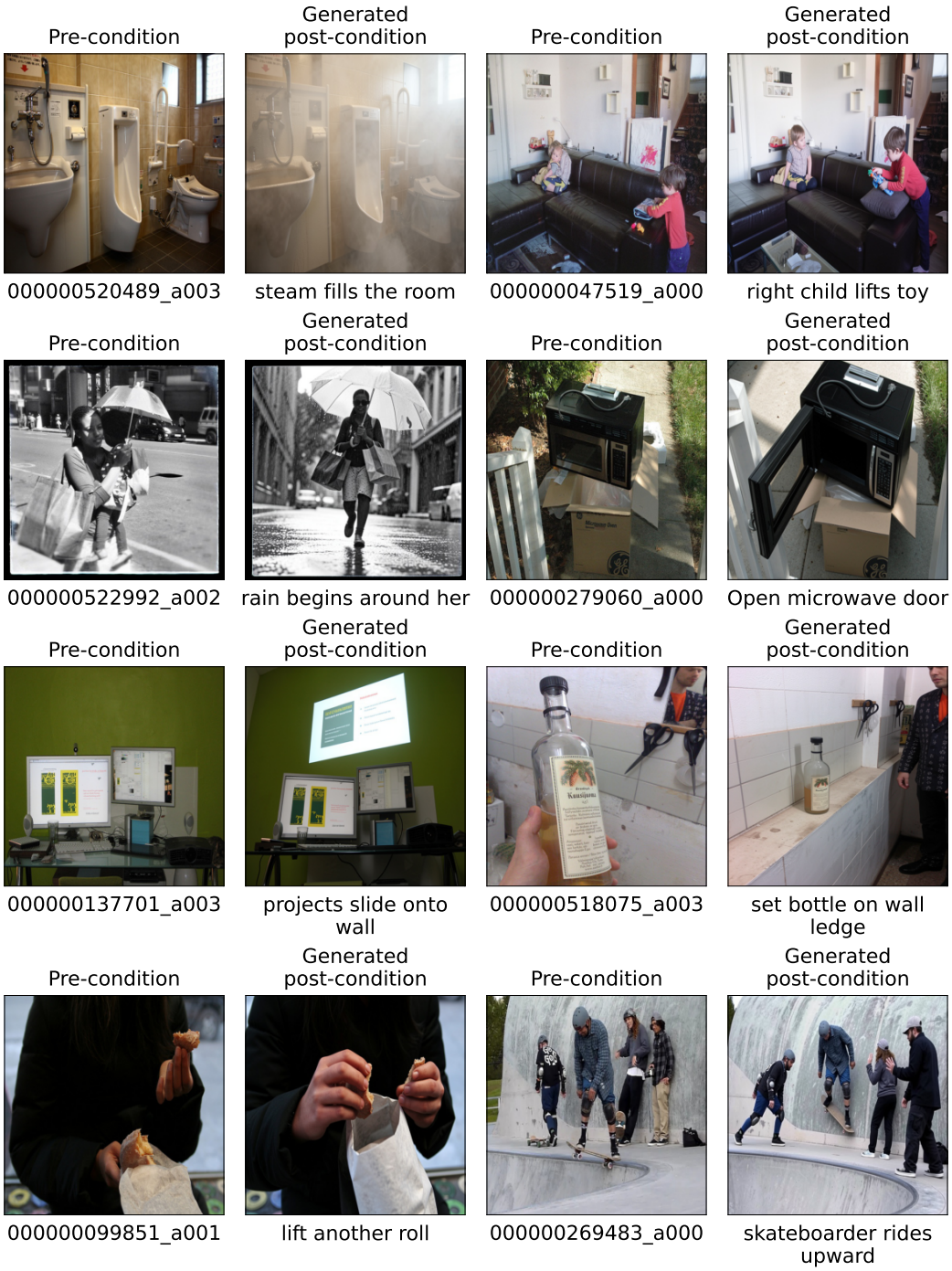



Figure 4: Sampled pre-/post-condition pairs in the world-model-distilled dataset. Each pair shows the source on the left and the generated state on the right.

instrument. The diagnostic permits the result to be left implicit. Result verbs specify a resulting change of state, location, or possession. Classification is a lemma-level lookup rather than a contextual semantic classifier. Consequently, the labels should be interpreted as the study's operational choices, especially for polysemous verbs.

The complete inventory contains 73 manner, 76 result, and 23 ambiguous entries.

**Manner (73).** bat, blow, braid, brush, carry, carve, cast, chop, clap, click, dip, draw, dribble, drink, drizzle, dunk, flick, fly, hit, hold, inspect, kick, knead, lick, mop, mow, offer, pass, pet, pitch, play, pour, press, puff, pull, push, reel, ride, rinse, rotate, rub, run, sand, scoop, scrape, scratch, serve, sew, shake, shave, shovel, signal, skip, slap, slide, smear, spin, spray, squeeze, step, stir, stretch, strike, sweep, swing, tap, throw, toss, touch, walk, wash, weld, wipe.

**Result (76).** add, adjust, align, appear, approach, arrange, assemble, attach, board, break, clean, close, come, complete, connect, cover, cross, cut, detach, disassemble, divide, drape, drop, dry, empty, enter, fill, flush, fog, fold, get, hang, insert, lead, light, lower, mark, mold, mount, open, pack, paste, peel, place, plant, plug, position, power, put, raise, reach, regain, release, remove, repair, retract, retrieve, secure, set, shift, shut, sit, slice, smooth, straighten, switch, take, tighten, tilt, transfer, trim, tuck, unfold, unscrew, water, withdraw.

**Ambiguous (23).** apply, attempt, catch, climb, dig, flip, grab, keep, lift, make, measure, mix, move, off, on, operate, paint, pick, roll, screw, shoot, tie, turn.

**Vocabulary counting.** Vocabulary sizes count distinct non-missing annotation labels within each split. Verb counts use the verb labels. Noun vocabularies use the object labels. Annotation strings are counted without additional normalization. Vocabulary sizes pooled across sources are computed from the union of their label sets.

# C Experiment

## C.1 Baselines and Ablation Variants

To reliably assess the contribution of each design choice in our framework, we conduct controlled ablation studies. We describe our baselines and variants below.

- $\mathrm{LM}_{\mathrm{fwd}} + \mathrm{Sim}_{\mathrm{bwd}} + \mathrm{Sim}_{\mathrm{fwd}}$: the main simulation-augmented setting, with 50% forward language modeling, 25% backward simulation, and 25% forward simulation. This tests word learning when action-text prediction is jointly trained with visual simulation in both directions.
- $\mathrm{LM}_{\mathrm{fwd}} + \mathrm{Sim}_{\mathrm{bwd}}$: removes forward simulation and keeps backward simulation. This isolates the contribution of inverse visual prediction.
- $\mathrm{LM}_{\mathrm{fwd}} + \mathrm{Sim}_{\mathrm{fwd}}$: removes backward simulation and keeps forward simulation. This isolates the contribution of future-state visual prediction.
- $\mathrm{LM}_{\mathrm{fwd}}$ only: removes both simulation objectives. This is the main language-only baseline, testing word learning from action-text prediction without image-embedding prediction.
- $\mathrm{LM}_{\mathrm{bwd}}$ only: uses backward language modeling without image-embedding prediction. This controls for whether reversing the order of the visual states changes the language-only baseline.
- $\mathrm{LM}_{\mathrm{fwd}} + \mathrm{LM}_{\mathrm{bwd}}$: uses both language-modeling directions without simulation. This tests whether exposure to both visual-state orders helps independently of image-embedding prediction.

For each variant, we train on each corpus with 3 random seeds for 10,000 steps. Unless otherwise noted, we treat full-sequence supervision with the InfoNCE image loss as our main setting and report contrasts with the cosine and completion-only variants only when they alter the qualitative picture.

## C.2 Preprocessing

Both corpora are converted to the same training format. All images are resized to $224 \times 224$ and normalized to $[-1, 1]$ before patch embedding. Each triple is rendered into one of the sequence formats defined in Section 2.2; the sampling probabilities over formats are determined by the training variant in Section C.1. Models differ only in the objective mixture, image loss, or supervision mask.

All variants use NEPA-large, with 24 transformer layers, a hidden dimension of 1,024, and 16 attention heads. The patch embedding, transformer, and final layer normalization are initialized from the same pretrained NEPA checkpoint. Word embeddings and one-dimensional positional embeddings are newly initialized, and the language-model output weights are tied to the word embeddings. Within each corpus, comparisons share the training split, vocabulary, preprocessing, and model architecture. Variants differ in sequence-format sampling, image loss, or supervision mask.