CONTINUAL VISUAL AND VERBAL LEARNING THROUGH A CHILD'S EGOCENTRIC INPUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Children learn the meanings of words from a continuous, temporally structured stream of egocentric experience. Recent work shows that neural networks can also learn word-referent mappings from a child's egocentric video recordings, but they cycle through the shuffled data for hundreds of epochs, contrasting with how children actually encounter their environment. We introduce BabyCL, a continual multimodal framework that combines streaming visual representation learning with an image–text contrastive objective to learn in a single chronological pass from naturalistic child-headcam video–language streams. BabyCL combines a multi-stage temporal segmentation of the stream with a dual replay buffer that independently manages visual and multimodal histories, and it jointly optimizes three objectives on a shared backbone. Against a streaming baseline that shares BabyCL's replay, schedule, and labeled-pair exposure but omits its visual objectives, BabyCL's frozen visual features yield consistently higher linear-probe accuracy, both in domain and on a separate child-headcam corpus. BabyCL also improves four-alternative forced-choice word–referent accuracy over this baseline and narrows the gap to offline training. Together, these results show that meaningful word-referent mappings can emerge under training conditions much closer to a child's actual experience.

1 INTRODUCTION

Well before their first birthday, children begin learning their first words, linking spoken words to their visual referents (Bergelson & Swingley, 2012). By age two, children are highly proficient word learners (Bloom, 2000): they can associate spoken words with objects, actions, and scenes that they encounter in everyday life. Children learn language despite a continuous stream of input that is noisy, temporally correlated, and highly redundant. Whether this natural input stream is sufficient for word learning or we need strong linguistic priors is a central question at the intersection of cognitive science and artificial intelligence.

One mechanism that could underlie this ability is *cross-situational word learning*, whereby children (or computational models) are sensitive to cross-situational regularities between visual and verbal input (Yu & Smith, 2007; Frank et al., 2007). For example, Vong & Lake (2022) showed that multimodal neural networks can also acquire word-referent mappings through cross-situational statistics, but this model and related models rely on curated laboratory stimuli or small-scale datasets. It is unclear whether such mechanisms scale to noisy, continuous, and temporally structured inputs that children actually receive. More recently, Vong et al. (2024) demonstrated that a contrastive image-text model akin to CLIP (Radford et al., 2021) trained on headcam video and child-directed speech from the SAYCam dataset (Sullivan et al., 2021) can acquire meaningful word-referent mappings. However, their training procedure processes the data in an offline, i.i.d. fashion with up to 400 cycles over the full dataset. This setup contrasts with the single-pass, temporally structured nature of human learning—a recent critique argues that this data cycling invalidates any cognitively-meaningful conclusions drawn from these models (Bowers, 2026).

Is the challenge of single-pass word learning addressable? There are some promising results on vision-only learning. Zhuang et al. (2021) established benchmarks for life-long visual learning on the SAYCam dataset (Sullivan et al., 2021) and evaluated a range of deep self-supervised algorithms, finding that contrastive methods that leverage negative sampling are most effective (e.g., SimCLR; Chen et al., 2020). Other work offers ways of alleviating the temporal redundancy issue through more sophisticated buffers (Purushwalkam et al., 2022). Closest to our work,

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.

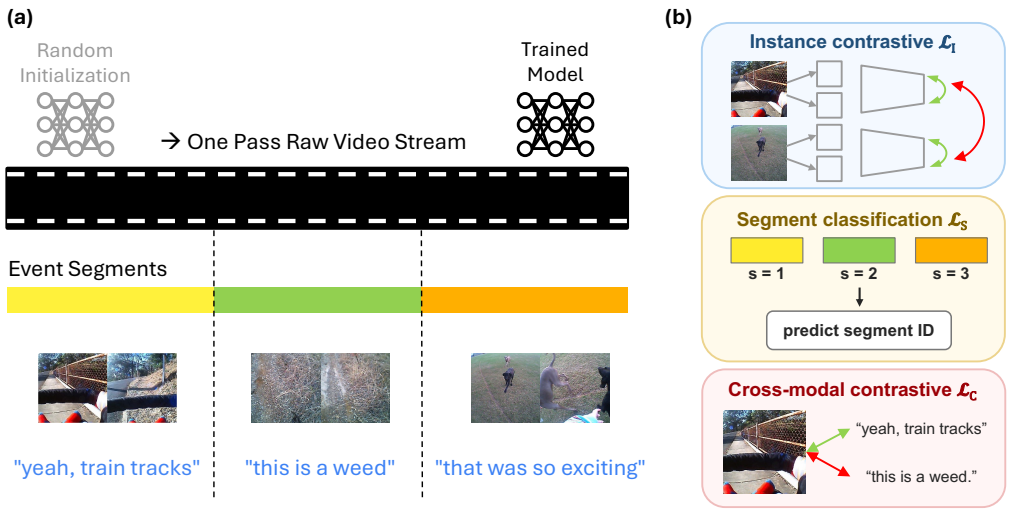


Figure 1: **Overview of BabyCL.** (a) The incoming SAYCam stream is partitioned into event segments of roughly three minutes using hierarchical visual clustering, with boundaries adjusted to utterance timestamps. (b) The visual branch applies instance-level contrastive learning to augmented frame views ($\mathcal{L}_I$) and predicts event-segment identities ($\mathcal{L}_S$), while the cross-modal branch contrasts matching utterance–frame pairs against mismatched pairs ($\mathcal{L}_C$).

Memory Storyboard (Yang & Ren, 2025) learns visual representations in a single pass over long-form egocentric video by grouping frames into temporal segments for replay. These methods address representation learning from dense visual streams, but not the asymmetric supervision of multimodal streams, where abundant video and sparse utterance–frame pairs must jointly shape a shared encoder. It therefore remains unclear whether replay alone is sufficient, or whether successful single-pass grounding requires temporal organization and complementary visual supervision.

We study multimodal representation learning from a temporally correlated video–language stream under a single-pass constraint, where dense visual observations and sparse linguistic supervision create distinct replay and optimization requirements. We introduce BabyCL, which organizes the stream into event-like segments, maintains separate visual and multimodal histories, and jointly learns instance-level, temporal, and cross-modal structure through a shared encoder. Unlike the contrastive vision-language model (CVCL) of Vong et al. (2024), whose CLIP objective uses only utterance-aligned frames, BabyCL applies instance-level and temporal objectives to the full video stream while using its cross-modal objective on aligned pairs. Together, chronological single-pass training, event-based organization, and bounded replay make BabyCL a closer computational approximation to learning from temporally structured experience than standard shuffled, multi-epoch training.

2 Related Work

Developmentally inspired learning. Several recent works have trained deep learning models on child-perspective sensory data, predominantly the SAYCam dataset (Sullivan et al., 2021). Orhan et al. (2020) showed the emergence of high-level visual category representations in self-supervised visual models trained on the temporal classification objective. Sheybani et al. (2023) demonstrated that the temporal ordering of infant experience provides a useful curriculum signal for visual representation learning. Yang & Ren (2025) further showed that with a memory buffer significantly smaller in size than the total training frames, self-supervised learning models can achieve competitive performance to offline training in a single-pass, streaming setting. These results suggest that self-supervised learning methods are highly effective for learning good visual representations from child-perspective visual data, but they remain limited to the visual modality.

Extending to grounded language learning, Vong et al. (2024) trained CVCL on egocentric video frames paired with transcribed child-directed speech from a single child in SAYCam, showing that the model acquires word-referent mappings and generalizes to novel referents. Vong & Lake (2026) established the robustness of these findings across all three children and multiple model architectures. BabyVLM (Wang et al., 2025a;b) explored generative vision-language models trained

on SAYCam data augmented with synthetic child-directed captions and introduced developmentally motivated evaluation benchmarks. These results provide empirical evidence that deep learning models can acquire word meanings from weakly supervised, naturally co-occurring vision-language pairs in naturalistic, continuous video streams. However, a key limitation is that these models were trained offline with shuffled data over multiple epochs, in contrast to how children experience their environment. Our work targets this limitation, raised by Bowers (2026) (Section 1), by learning grounded multimodal representations from a single temporally ordered pass, substantially reducing reliance on shuffled, multi-epoch training.

Continual representation learning. Continual Representation Learning (Rao et al., 2019; Fini et al., 2022; Gomez-Villa et al., 2022; Zhang et al., 2024) addresses the problem of learning good representations from unlabeled, non-stationary data streams. Purushwalkam et al. (2022) identified temporal correlation and distribution shift as key challenges for continuous self-supervised learning and proposed minimum-redundancy (MinRed) replay buffers to mitigate these issues. Madaan et al. (2022) demonstrated that unsupervised representations are more robust to catastrophic forgetting than supervised ones and mixing current and past instances further alleviates forgetting. Recent work has also begun to address continual learning for vision-language models (Liu et al., 2025b;a). However, these works focus on adapting pretrained models to shifting task distributions, rather than learning multimodal representations from scratch in a developmentally motivated streaming setting. Our work instead combines instance-level visual contrastive learning and temporal classification with an image–text contrastive objective to learn a joint image–text representation through child-centric data in a streaming setting. Unlike task-based continual learning, our setting has no task boundaries or task labels and learns representations from scratch. It is also not a few-shot or small-data setting; the challenge arises from a single chronologically ordered, highly correlated stream.

3 METHOD

We present **BabyCL**, a framework for continual grounded language learning from a continuous video stream paired with aligned text. The pipeline is illustrated in Figure 1 and performs visual and verbal learning simultaneously under a shared backbone: 1) the streaming visual representation learning framework of Yang & Ren (2025), which uses hierarchical temporal segmentation and a two-tier replay buffer, and 2) the CVCL image-text contrastive objective of Vong et al. (2024).

Temporal segmentation. Following Yang & Ren (2025), we partition the incoming stream into event segments of roughly three minutes by choosing change points that maximize the within-segment cosine similarity of frame embeddings (Appendix A). To accommodate the online streaming learning setting, we apply temporal segmentation in the short-term memory. Segmentation is re-run every 12 s of incoming video, interleaved with training, and an event segment enters the short-term memory once it ends at least 12 s before the newest frame; segments already in memory are never revised. We further adapt this temporal segmentation algorithm to our multimodal setting by constraining the change points to fall outside any utterance interval, enforcing that every utterance from the transcript lies within a single event segment.

Memory replay. To mitigate temporal redundancy and catastrophic forgetting in the single-pass stream, BabyCL maintains two replay buffers. The visual buffer stores past event segments for frame-level self-supervision. The multimodal buffer stores utterance-frame pairs; following Vong et al. (2024), a frame is randomly sampled from within each utterance's timestamp window. Each buffer is split into a short-term first-in, first-out (FIFO) tier and a long-term reservoir-sampled tier, as in Yang & Ren (2025). At each step, the default training batch contains 96 visual-buffer frames, organized as three 32-frame event segments, and 32 utterance–frame pairs from the multimodal buffer, corresponding to a 75%/25% split. Within each buffer, 25% of the samples are drawn from the short tier and 75% from the long tier. For the experiments reported here, the short and long tiers hold by default 10% and 50% of the total stream, respectively (hyperparameters we vary in Section 4.5).

Resampling at insertion. Because SAYCam is traversed in a single chronological pass, the total number of gradient steps is bounded by the length of the stream. To ensure that training converges within this budget, each time the newly finalized event segments are inserted into the visual buffer, we perform k total forward-backward passes (8 by default) on batches that mix the newly inserted



Figure 2: **Example of the 4AFC task.** Given a target category ("ball") and three foil categories ("cat", "couch", and "car"), the model must select the image corresponding to the target category from four candidate images.

Table 1: **Labeled-S 4AFC image-mode accuracy.** Mean ± approximate 95% CI over five seeds; chance is 25%. Offline rows train on shuffled data for multiple epochs, whereas the rows below the rule learn in a single chronological pass with k updates per buffer insertion.

Method	4AFC (%)
Offline CVCL (i.i.d.)	57.31 ± 1.13
Offline CVCL (compute-matched)	46.54 ± 1.97
CL-CVCL ($k = 1$)	31.48 ± 3.74
CL-CVCL ($k = 8$)	40.72 ± 1.49
BabyCL ($k = 4$)	41.08 ± 1.79
BabyCL ($k = 8$)	**43.38 ± 1.47**

material with replayed content. The chronological order in which new segments first enter the stream is preserved; resampling only affects how often past content is revisited during training.

Training objectives. All samples are encoded by a shared ResNeXt-50 visual backbone (matching CVCL (Vong et al., 2024)), with modality-specific projection heads. BabyCL is trained with three losses applied jointly at every step:

- **Instance loss** $\mathcal{L}_I$, a SimCLR objective (Chen et al., 2020) on augmented views of frames sampled from the visual buffer.
- **Temporal loss** $\mathcal{L}_S$, a supervised classification loss over event segments (Orhan et al., 2020; Yang & Ren, 2025). Labels are batch-local: the $S = 3$ event segments in each visual minibatch are indexed $0, \ldots, S-1$ and each frame is classified into its own segment, so the classifier head has a fixed size and does not grow with the stream.
- **Cross-modal loss** $\mathcal{L}_C$, a symmetric InfoNCE loss on utterance-frame pairs drawn from the multimodal buffer, matching CVCL (Vong et al., 2024).

The three losses are combined into a single objective:

$$\mathcal{L} = 0.5\,(\mathcal{L}_I + \mathcal{L}_S) + 0.5\,\mathcal{L}_C. \tag{1}$$

4 Experiments

4.1 Datasets and evaluation

SAYCam. Our primary experiments use the child S subset of SAYCam (Sullivan et al., 2021), which contains 221 hours of naturalistic video collected from a head-mounted camera on the child from age 6–32 months, decoded at 5 fps. For multimodal supervision, we use the transcribed child-directed speech curated by Vong et al. (2024), comprising 37,486 utterances paired with 600,285 image frames. We train on this stream in chronological order and evaluate on Labeled-S.

BabyView. We additionally consider BabyView (Long et al., 2024), a more recent corpus of infant egocentric video with automatically generated transcripts. The stream we use comprises 409 hours of video (7.36M frames at 5 fps) from 32 children, paired with 259,721 transcribed utterances; the full corpus spans ages from 6 months to 3 years (Long et al., 2024). We use BabyView in two complementary ways: first, to test whether models trained on SAYCam transfer to an independent child-egocentric corpus; and second, to test whether BabyCL can learn directly from a different naturalistic video-language stream.

Evaluation. We report accuracy on the four-alternative forced-choice (4AFC) task on Labeled-S (Orhan et al., 2020), following Vong et al. (2024); Vong & Lake (2026), shown in Figure 2. Each trial presents one target category and three foils; the model scores four candidate images against the target category name and must identify the matching one. We introduce the corresponding BabyView evaluation below. As the official manually screened trial file was unavailable to us, we regenerate trials with the same procedure (Section 4.2).

Table 2: **BabyView zero-shot 4AFC.** Accuracy (%) on 4,150 frozen trials over the 61 in-vocabulary categories; chance 25%. Each model uses the end-of-training checkpoint of its best seed; BabyCL's five-seed mean is 33.37% (Table 4).

Method	4AFC (%)
Offline CVCL	35.18
CL-CVCL ($k = 1$)	29.06
CL-CVCL ($k = 8$)	29.16
BabyCL	**32.53**

Table 3: **BabyView linear probe.** 72-way top-1 accuracy (%) on frozen pooled backbone features; chance 1.4%. Each model uses the end-of-training checkpoint of its best seed.

Method	Top-1 (%)
Offline CVCL	57.37
CL-CVCL ($k = 1$)	22.75
CL-CVCL ($k = 8$)	35.72
BabyCL	**49.96**

4.2 Results on SAYCam

We compare BabyCL to four CVCL-style baselines that isolate different explanations for streaming performance. **Offline CVCL** (Vong et al., 2024) trains on i.i.d. shuffled data; we report both its full 400-epoch result, which serves as an upper reference point for the CVCL objective, and a compute-matched result at 25 epochs, which matches the number of labeled utterance–frame pairs seen by BabyCL (Appendix A). **CL-CVCL** is the direct ablation of BabyCL's visual objectives: it removes $\mathcal{L}_I$ and $\mathcal{L}_S$ (their coefficients set to zero) while keeping the stream, segmentation, replay buffers, batch composition (32 labeled pairs per step), optimizer, learning-rate schedule, input normalization, and step count identical, so hyperparameters and labeled-pair exposure are matched by construction. We run it with $k = 1$ (one update per insertion) and with BabyCL's default $k = 8$ (compute-matched). Table 1 reports 4AFC accuracy. CL-CVCL with a single update per insertion reaches only 31.48%; with BabyCL's full update budget ($k = 8$) it reaches 40.72%, so revisiting replayed pairs accounts for most of the streaming gain. Because CL-CVCL at $k = 8$ differs from BabyCL only in the visual objectives, BabyCL's further gain to 43.38% is attributable to $\mathcal{L}_I$ and $\mathcal{L}_S$; on 4AFC this gain is modest (about 2.7 points). Offline CVCL trained i.i.d. on the same number of labeled pairs as BabyCL reaches 46.54%: BabyCL's advantage is over streaming training under a matched budget, not over i.i.d. training with equal pair exposure. Increasing the total number of insertion-time updates from $k = 4$ to $k = 8$ yields a further 2.3-point gain, showing that revisiting newly inserted content more thoroughly adds to the benefit of the continual-learning structure. BabyCL substantially closes the gap to offline CVCL (57.31%) but does not match it. Our offline CVCL reaches 57.31% rather than the ∼61.6% of Vong et al. (2024) because our regenerated trials lack the official set's manual screening and are therefore harder; all models are evaluated on the same trials.

4.3 Transfer from SAYCam to BabyView

To evaluate models on the second corpus, we construct a new BabyView 4AFC benchmark from 5,921 public, rater-validated object crops. It mirrors the Labeled-S protocol with 4,150 fixed trials over the 61 BabyView categories present in the SAYCam vocabulary, but uses independently collected crops with varied framing and aspect ratios. We additionally conduct a 72-way linear probe of frozen visual features. These tasks provide a challenging test of cross-corpus transfer rather than a difficulty-matched replacement for Labeled-S. Appendix G details the benchmark construction. As Tables 2 and 3 show, BabyCL leads the streaming models on both tasks, and the ranking matches Labeled-S. The corpus shift narrows the offline–streaming 4AFC gap from about 14 points in domain to 2.7 (35.18% vs. 32.53%), and BabyCL's probe advantage over CL-CVCL ($k = 8$) (49.96% vs. 35.72%) indicates that its visual objectives yield features that transfer across child-egocentric corpora.

4.4 Training on SAYCam and BabyView

We also train BabyCL directly on the BabyView stream with the default recipe, changing only corpus-specific data handling (a fixed aspect-ratio pre-crop for portrait frames, utterance text normalized to the SAYCam vocabulary, and a segmentation window scaled to the longer stream); each configuration uses five seeds and end-of-training checkpoints. As Table 4 shows, the BabyView-trained model is lower on both Labeled-S (28.54% vs. 43.38%) and BabyView itself (28.38% vs. 33.37%), likely reflecting noisier BabyView supervision and a recipe not retuned for the new corpus; we leave such tuning to future work. Appendix H addresses the supervision noise by weighting utterance–frame pairs according to a credibility score computed with a pretrained CLIP ViT-B/32; since this departs

Table 4: **Training corpus × evaluation set.** End-of-training 4AFC accuracy of the default BabyCL recipe trained on each corpus (rows) and evaluated on each benchmark (columns); mean ± approximate 95% CI over five seeds.

	Evaluation set (4AFC, %)	
Training corpus	Labeled-S	BabyView
SAYCam	43.38 ± 1.47	33.37 ± 0.60
BabyView	28.54 ± 1.79	28.38 ± 3.53

Table 5: **Contribution of the visual objectives.** Labeled-S 4AFC accuracy when one or both visual objectives are removed from the default BabyCL recipe; removing both yields CL-CVCL ($k = 8$). Results are mean ± approximate 95% CI over five seeds.

Training objectives	4AFC (%)
$\mathcal{L}_C + \mathcal{L}_I + \mathcal{L}_S$	**43.38 ± 1.47**
$\mathcal{L}_C + \mathcal{L}_I$	42.30 ± 0.86
$\mathcal{L}_C + \mathcal{L}_S$	40.79 ± 1.75
$\mathcal{L}_C$ (CL-CVCL)	40.72 ± 1.49

Table 6: **Effect of visual buffer size.** We vary the short-term and long-term visual buffer capacities as a percentage of the total video stream. Reported values show mean accuracy with approximate 95% confidence intervals over five seeds. CL-CVCL uses $k = 8$.

Method	Short buffer	Long buffer	4AFC (%)
CL-CVCL	10%	50%	40.72 ± 1.49
BabyCL	2%	10%	33.78 ± 1.96
BabyCL	5%	25%	39.72 ± 1.57
BabyCL	10%	50%	**43.38 ± 1.47**

from our from-scratch setting, we treat it as a separate analysis, and no other result in the paper uses it. Table 13 shows the resulting gains.

4.5 Ablation experiments and other training factors

Contribution of the visual objectives. To measure the contribution of the two visual objectives, we train variants that remove either $\mathcal{L}_S$ or $\mathcal{L}_I$ from the default configuration. The remaining loss coefficients, cross-modal objective, and all other hyperparameters are unchanged. Each variant is run with five random seeds; the default row is the BabyCL ($k = 8$) result from the main experiments.

Table 5 shows that $\mathcal{L}_I$ carries most of the benefit: $\mathcal{L}_C + \mathcal{L}_I$ (42.30%) remains above CL-CVCL (40.72%), whereas $\mathcal{L}_C + \mathcal{L}_S$ (40.79%) is comparable to it. $\mathcal{L}_S$ nevertheless adds to $\mathcal{L}_I$: the full combination has the highest point estimate (43.38%), suggesting that the two objectives are complementary, although these differences lie within the confidence intervals.

Balancing visual and language-grounded supervision. We investigate how different proportions of language-grounded examples in each batch affect training performance, and summarize the outcomes in Table 7. Specifically, we vary the fraction of examples drawn from the multimodal buffer in each batch among 25%, 50%, and 75%, with the remainder drawn from the visual buffer.

The 50% mixture has the highest final point estimate, but the differences among the three regimes are small relative to their uncertainty. Performance is therefore broadly stable across the tested mixtures.

Robustness to replay buffer design. We investigate how sensitive BabyCL is to the buffer management policy, keeping all other training choices fixed. The same policy is applied to both the video and the multimodal buffers, with the 10%/50% short/long capacity split used in the main runs. We compare the default two-tier *FIFO + reservoir* policy (short-tier + long-tier eviction) against *FIFO + FIFO*, *FIFO + random*, and a *single-tier reservoir* that merges both tiers into a flat reservoir of equal total capacity.

As Table 8 shows, the four policies fall within roughly 1.4 percentage points of each other (41.98%–43.38%), with the default *FIFO + reservoir* configuration performing best. The two-tier variants stay within about one point of the default, so BabyCL is fairly robust to the specific long-tier eviction rule; the *single-tier reservoir*, which removes the dedicated channel for recent items, trails by the largest margin.

Size of the visual buffer. The default visual replay buffer in BabyCL is relatively large: the short-term and long-term tiers store examples equivalent to 10% and 50% of the full video stream, respectively. We acknowledge that this memory budget is substantial for long streams, and therefore evaluate whether BabyCL remains effective under smaller visual replay buffers. Specifically, we

Table 7: **Balancing visual and language-grounded supervision.** Labeled-S 4AFC when multimodal (MM) pairs occupy 25%, 50%, or 75% of each minibatch (n=5 seeds). *Final mean* $\pm$ *CI* is the mean end-of-training accuracy with an approximate 95% confidence interval.

MM frac.	Final mean $\pm$ CI (%)
25%	43.38 $\pm$ 1.47
50%	**43.43 $\pm$ 1.36**
75%	42.67 $\pm$ 1.29

Table 8: **Effect of replay buffer policy on Labeled-S 4AFC accuracy.** The same policy is applied to both buffers. Reported values show mean accuracy with approximate 95% confidence intervals over five seeds. "Random" evicts a uniformly chosen long-tier resident on overflow.

Policy	4AFC (%)
FIFO + reservoir (default)	**43.38 $\pm$ 1.47**
FIFO + FIFO	43.07 $\pm$ 1.84
FIFO + random	42.49 $\pm$ 2.96
Single-tier reservoir	41.98 $\pm$ 2.79

Table 9: **Backbone ablation.** Labeled-S 4AFC accuracy (mean $\pm$ approximate 95% CI over five seeds) when the vision backbone is replaced under the fixed BabyCL training recipe. All backbones are trained from scratch; each uses its own backbone learning rate and weight decay.

Architecture	Backbone LR	Weight decay	4AFC (%)
ResNet-18	10^{-3}	10^{-4}	42.88 $\pm$ 1.16
ResNeXt-50 (default)	10^{-3}	10^{-4}	**43.38 $\pm$ 1.47**
ViT-S/16	3×10^{-4}	0.05	42.36 $\pm$ 1.79
ViT-Tiny	5×10^{-4}	0.05	42.17 $\pm$ 1.92
ViT-B/14	10^{-4}	0.05	39.04 $\pm$ 1.05
ConvNeXt-Tiny	10^{-4}	0.05	36.76 $\pm$ 1.22

reduce the short-term/long-term buffer capacities to 5%/25% and 2%/10% of the stream while keeping the rest of the training recipe fixed. Results are shown in Table 6.

We observe that reducing the visual buffer size leads to a clear drop in downstream performance, indicating that replay capacity is important for maintaining diverse visual experience from the stream. BabyCL's advantage over the compute-matched CL-CVCL baseline (40.72%) requires the default buffer: with a halved 5%/25% buffer, BabyCL (39.72%) falls slightly below it, and with the smallest 2%/10% buffer (33.78%) well below it, so a sufficient replay budget is a prerequisite for the benefit of the visual objectives.

Vision backbone. We replace ResNeXt-50 with other architectures trained from scratch, each with its own backbone learning rate and weight decay (Appendix D). As Table 9 shows, ResNet-18, ViT-S/16, and ViT-Tiny fall within about one point of ResNeXt-50, so BabyCL does not depend on a particular encoder. ViT-B/14 and ConvNeXt-Tiny trail behind, likely because a single pass provides too few updates for these data-hungry architectures.

4.6 Linear probing without finetuning

Forced-choice alignment scores measure vision–language alignment rather than the discriminability of the visual features. To assess the visual features alone, we train linear classifiers on frozen backbone embeddings. We fit multinomial logistic-regression probes on frozen image embeddings for Labeled-S (26-way) and mini-ImageNet (100-way); see Appendix B.

BabyCL gains +4.84 points over CL-CVCL ($k = 8$) on the Labeled-S probe and +12.28 on mini-ImageNet (Table 10): the visual objectives improve the frozen visual representation itself, most clearly on images outside the training distribution.

4.7 Visual Two-Word Test (VTWT) and Baby Winoground

We further evaluate all models zero-shot on Visual Two-Word Test (VTWT) and Baby Winoground, two in-domain evaluation benchmarks proposed by Wang et al. (2025a) for multimodal models trained on SAYCam. Scoring details are given in Appendix C.

Visual Two-Word Test (VTWT). VTWT probes compositionality by asking the model to match SAYCam frames with appropriate two-word phrases (e.g., "wash cup" vs. "fill cup"). Table 10 reports the results. All models score above chance (50%), and BabyCL comes within about 2 points of

Table 10: **Additional evaluations.** Accuracy (%). *Linear probe*: top-1 accuracy of frozen visual embeddings on Labeled-S (26-way classification; per-class 80/20 train/test split) and mini-ImageNet (100-way). *VTWT* and *Baby Winoground* are zero-shot with no additional training; Baby Winoground *Overall* requires both the Positive Ctx (original scene) and Negative Ctx (edited scene) checks to hold. Each model uses the end-of-training checkpoint of its best seed. Bold marks the best streaming method; Offline CVCL is a non-streaming reference.

	Linear probe			Baby Winoground		
Method	Labeled-S	mini-ImageNet	VTWT	Overall	Pos. Ctx	Neg. Ctx
Chance	*3.85*	*1.00*	*50.00*	*25.00*	*50.00*	*50.00*
Offline CVCL	60.17	37.22	63.08	11.51	63.01	41.92
CL-CVCL ($k = 1$)	24.82	4.72	55.43	7.12	57.53	44.11
CL-CVCL ($k = 8$)	40.11	13.32	58.63	12.05	**60.82**	43.84
BabyCL	**44.95**	**25.60**	**61.32**	**14.79**	59.73	**46.85**

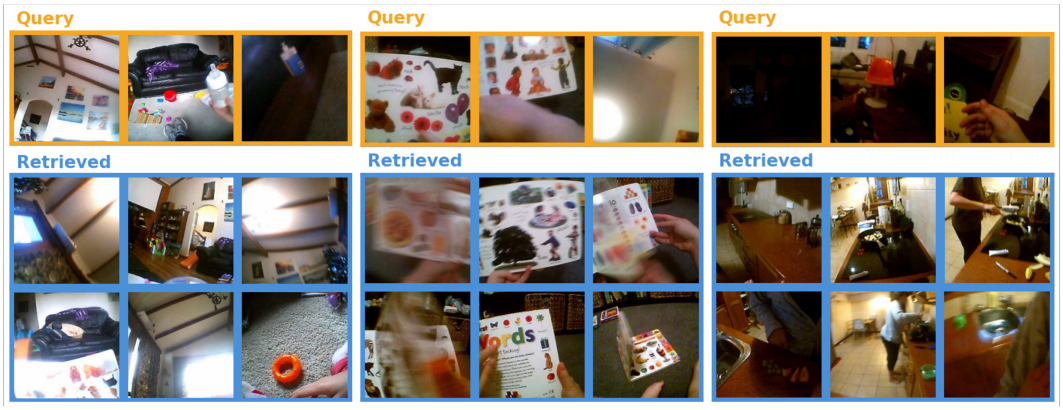


Figure 3: **Cross-video nearest neighbors (BabyCL).** Each column is one query segment (top, yellow) and its retrieved segments from other recordings (below, blue).

Offline CVCL. Increasing the number of updates per insertion from $k = 1$ to $k = 8$ raises CL-CVCL from 55.43% to 58.63%, and the visual objectives add a further 2.69 points (61.32%).

Baby Winoground. Baby Winoground couples two simultaneously observed scenes with the same contrasting phrase pair such that each image warrants exactly one wording under the intended semantics. The benchmark evaluates whether a model can preserve correct image–text associations under both the original SAYCam context and a synthetically perturbed negative context. An item counts as correct (*Overall*) only if the original scene prefers its faithful caption (*Positive Ctx*) *and* the edited scene prefers the alternate phrase (*Negative Ctx*).

Table 10 summarizes the results. Offline CVCL achieves the highest Positive Context score (63.01%), while BabyCL has the highest Negative Context (46.85%) and Overall (14.79%) scores, compared with 11.51%, 12.05%, and 7.12% Overall for Offline CVCL, CL-CVCL ($k = 8$), and CL-CVCL ($k = 1$). We treat this result as a complementary, hypothesis-generating analysis rather than evidence of strong compositional understanding or general robustness to context shifts. One possible interpretation is that temporally structured training encourages more consistent paired rankings, but establishing this mechanism requires dedicated controls and uncertainty estimates.

4.8 Qualitative Analysis

Nearest-neighbor retrieval. For each event segment (Yang & Ren, 2025) we average 512-dimensional image embeddings over sampled frames, ℓ_2-normalize, and rank all other segments by cosine similarity, excluding the query recording itself.

Figure 3 retrieves living-room play, picture-book reading, and kitchen scenes from recordings disjoint from each query, i.e. recurring scene types rather than temporally adjacent frames of the same recording.

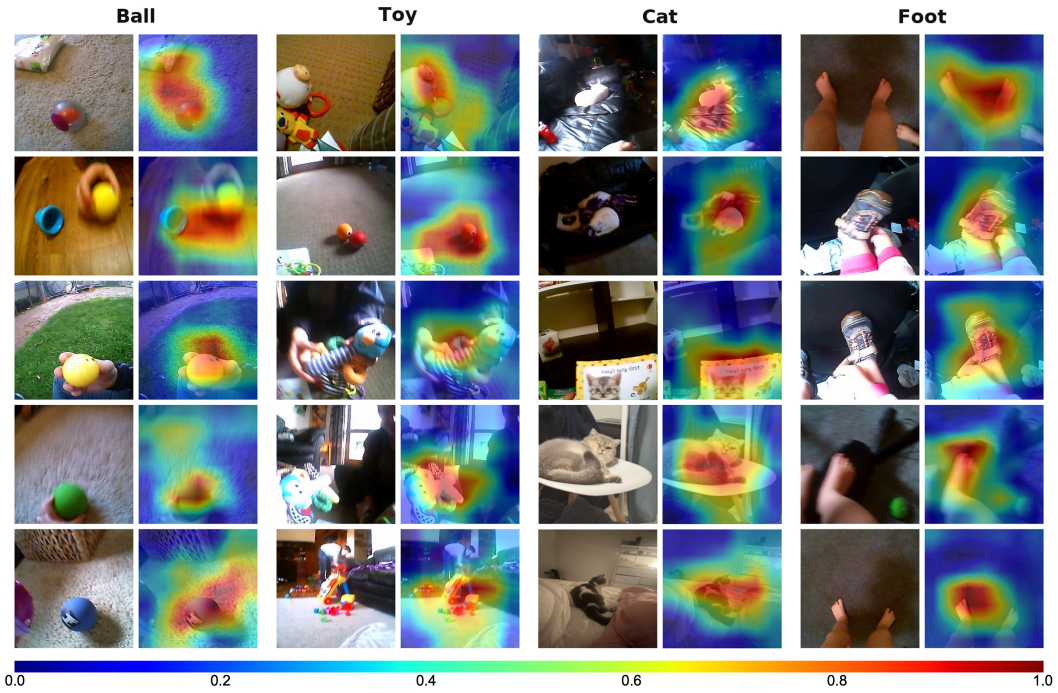


Figure 4: **Grad-CAM++ with noun queries (BabyCL).** Rows: ball, toy, cat, foot; original frame above its overlay. Frames are those BabyCL scores most confidently for each noun, i.e. selected rather than random.

Grad-CAM++ localization. Grad-CAM++ (Chattopadhay et al., 2018) localizes where the image–text cosine alignment peaks inside each frame. For several Labeled-S nouns we retain infant-view exemplars that BabyCL scores most confidently against the embedded category token, backpropagate that cosine through the deepest convolutional maps of the frozen backbone, and overlay upsampled attributions at full resolution.

Figure 4 concentrates activation on balls, small toys, cats, and feet while most of the cluttered domestic background stays inactive, consistent with the cross-modal objective $\mathcal{L}_C$ tying image regions to specific nouns rather than to the scene as a whole. As these are BabyCL's most confidently scored frames, they show favorable rather than typical cases.

5 Discussion and Conclusion

BabyCL learns word–referent mappings from a single chronological pass over a child's video–language stream, without shuffling or epochs and with a bounded replay budget. Revisiting replayed pairs accounts for most of the gain over naive streaming, and the visual objectives, applied to the dense unlabeled stream, add a modest further improvement in word–referent accuracy. Their clearer benefit lies in the visual representation itself: frozen features improve in domain and more strongly on out-of-domain images and a separate child-headcam corpus. Offline training over many shuffled epochs, and even i.i.d. training with equal pair exposure, remain stronger; closing this gap under the streaming constraint is the main open problem.

Limitations and future work. A notable performance gap remains between BabyCL and offline CVCL; future work should aim to close it further. The replay buffers used here are large relative to the size of the video stream. Finally, additional evaluation benchmarks probing compositionality and generalization would give a more comprehensive picture.

AI use statement

In this work, we used generative AI tools for designing and providing feedback on research methodology and experiments, implementing methods, assisting with translation, and interpreting results. We

have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, cleaning and reformatting datasets, or supporting qualitative and thematic data analysis. We have reviewed all AI-assisted work: AI-generated code was inspected and tested by the authors before use, reported numbers were checked against the raw evaluation outputs, and AI-translated text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work uses naturalistic egocentric video and speech recorded in the homes of young children. SAYCam (Sullivan et al., 2021) and BabyView (Long et al., 2024) were collected by their original investigators with informed parental consent, and both are distributed through Databrary under authorized-access and data-use agreements, which we followed. We did not collect any new data from human participants, and beyond a small number of illustrative frames in the figures, we do not redistribute raw video, audio, transcripts, or frames from either corpus. The BabyView evaluation benchmark is constructed only from publicly released, rater-validated object crops. Our models are small, trained from scratch on child egocentric corpora, and are not designed to identify, track, or profile individuals; we therefore see limited potential for direct misuse.

Reproducibility statement

Appendix A specifies the architecture, objectives, optimizer and schedule, augmentation, online segmentation, replay composition, and update and labeled-pair budgets of BabyCL and all baselines. Evaluation protocols are described in Appendices B, C, and G, backbone-specific settings in Appendix D, and compute resources in Appendix I. SAYCam and BabyView are available to authorized researchers through Databrary.

References

Elika Bergelson and Daniel Swingley. At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, 2012.

Paul Bloom. *How children learn the meanings of words*. MIT Press, 2000. doi: 10.7551/mitpress/3577.001.0001.

Jeffrey S. Bowers. The successes and failures of artificial neural networks (ANNs) highlight the importance of innate linguistic priors for human language acquisition. *Psychological Review*, 133 (4):994–1005, 2026. doi: 10.1037/rev0000595.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 839–847. IEEE, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9621–9630, 2022.

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*, volume 20, pp. 457–464, 2007.

Alex Gomez-Villa, Bartlomiej Twardowski, Lu Yu, Andrew D. Bagdanov, and Joost van de Weijer. Continually learning self-supervised representations with projected functional regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3866–3876, 2022. doi: 10.1109/CVPRW56347.2022.00432.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-clip: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 53971–54012, 2025b.

Bria Long, Robert Z. Sparks, Violet Xiang, Stefan Stojanov, Zi Yin, Grace E. Keene, Alvin W. M. Tan, Steven Y. Feng, Chengxu Zhuang, Virginia A. Marchman, Daniel L. K. Yamins, and Michael C. Frank. The babyview dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *arXiv preprint arXiv:2406.10447*, 2024.

Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *The Tenth International Conference on Learning Representations*, 2022.

Emin Orhan, Vaibhav Gupta, and Brenden M Lake. Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, 33:9960–9971, 2020.

Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. In *European Conference on Computer Vision*, pp. 702–721. Springer, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in neural information processing systems*, 32, 2019.

Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36: 54199–54212, 2023.

Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C. Frank. SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, 5:20–29, 2021. doi: 10.1162/opmi_a_00039.

Wai Keen Vong and Brenden M Lake. Cross-situational word learning with multimodal neural networks. *Cognitive science*, 46(4):e13122, 2022.

Wai Keen Vong and Brenden M. Lake. On the robustness of modeling grounded word learning through a child's egocentric input. *Open Mind*, 10:1253–1278, 2026. doi: 10.1162/opmi.a.377.

Wai Keen Vong, Wentao Wang, A Emin Orhan, and Brenden M Lake. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511, 2024.

Shengao Wang, Arjun Chandra, Aoming Liu, Venkatesh Saligrama, and Boqing Gong. Babyvlm: Data-efficient pretraining of vlms inspired by infant learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1380–1390, 2025a.

Shengao Wang, Wenqi Wang, Zecheng Wang, Max Whitton, Michael Wakeham, Arjun Chandra, Joey Huang, Pengyue Zhu, Helen Chen, David Li, et al. Babyvlm-v2: Toward developmentally grounded pretraining and benchmarking of vision foundation models. *arXiv preprint arXiv:2512.10932*, 2025b.

Yanlai Yang and Mengye Ren. Memory storyboard: Leveraging temporal segmentation for streaming self-supervised learning from egocentric videos. *arXiv preprint arXiv:2501.12254*, 2025.

Chen Yu and Linda B Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, 18(5):414–420, 2007.

Yipeng Zhang, Laurent Charlin, Richard Zemel, and Mengye Ren. Integrating present and past in unsupervised continual learning. *arXiv preprint arXiv:2404.19132*, 2024.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

A Implementation details

This appendix records concrete architectural and optimization details for BabyCL and the CVCL-family baselines.

Encoders and projection layers. The trained vision backbone of every model is a ResNeXt-50 initialized from scratch. The only pretrained networks are kept frozen: an ImageNet-pretrained ResNeXt-50 for temporal segmentation, CLIP ViT-B/32 for credibility weighting (Appendix H), and the CLIP ViT-B/16 text encoder in the frozen-teacher reference experiment (Appendix F). We replace the classifier head with a single linear layer that maps the final ResNeXt block output (2048-d) to a 512-d embedding. The text side uses a learned 512-d word embedding table; an utterance embedding is the mean of its token embeddings (with out-of-vocabulary tokens mapped to the unknown-token index).

BabyCL: streaming visual branch. The visual branch uses a SimCLR-style objective with a multilayer perceptron (MLP) projector of sizes $[d, 2048, 128]$ (ReLU between layers), where d is the backbone feature dimension (2048 for ResNeXt-50). NT-Xent on paired augmented views uses temperature 0.2. The temporal classification component uses a second MLP of sizes $[d, 2048, C]$ with a fixed $C = 2000$ outputs; because labels are batch-local, only the first $S = 3$ outputs ever receive targets.

BabyCL: multimodal branch. The multimodal branch encodes images with the same vision backbone (and the 2048→512 projection) and encodes text with the 512-d mean-pooled token embeddings. Image and text embeddings are ℓ_2-normalized; we optimize symmetric image–text InfoNCE at temperature 0.07 (distinct from the visual NT-Xent temperature). Replay mixes short- and long-term buffers without semantic similarity filtering.

BabyCL: optimizer and schedule. We optimize with AdamW under automatic mixed precision, using the usual moment hyperparameters (β_1=0.9, β_2=0.999). Four learning-rate tiers apply: 10^{-3} for the shared ResNeXt convolutional stack, the embedding head that maps pooled features to the 512-dimensional image embedding, and other decaying weights in that encoder (batch-normalization affine parameters and biases are trained at the same rate but omit weight decay), 10^{-4} for the SimCLR and temporal-classification heads built on the 2048-dimensional backbone output in the streaming branch, 10^{-4} for multimodal modules outside that encoder, and 10^{-4} for the small language-model stack excluding the shared token embedding table. Weight decay is 10^{-4} wherever applied; scale and bias parameters in batch-normalization layers (and analogous no-decay terms in the encoder) use zero weight decay. After a linear warmup lasting 2% of the scheduled optimizer steps (at least 100 steps), every tier follows the same cosine multiplier that decays to zero by the end of training. The main online runs use batch size 128.

BabyCL: data augmentation. For self-supervised learning (SSL), we generate two views per frame using strong augmentation: random resized crop to 224×224 (scale 0.2–1.0), random horizontal flip, color jitter (0.8/0.8/0.8/0.2 applied with 0.8 probability), random grayscale (0.2), Gaussian blur (applied with probability 1.0 on view 1 and 0.1 on view 2), and solarization on view 2 with probability 0.2. Inputs are normalized with ImageNet channel statistics.

BabyCL: online segmentation and replay composition. Temporal segmentation embeds one frame every 60 frames (12 s) with f_θ, a frozen ImageNet-pretrained ResNeXt-50 from torchvision that is used only for segmentation and is separate from the model being trained, and scores a segment s by $\frac{1}{|s|}\sum_{i,j\in s}\cos\big(f_\theta(x_i), f_\theta(x_j)\big)$, following Yang & Ren (2025). Each utterance, located by its timestamps, forms its own piece and is never split. Each unannotated gap of length M between utterances is split into $\max(1, \mathrm{round}(M/T))$ pieces, where the hyperparameter T is set to 900 frames (three minutes at 5 fps); change points are added greedily one at a time, each maximizing the total

segment score. The pieces are then packed in temporal order into event segments, each closed once it spans at least T frames. Each step trains on 128 frames in four 32-frame segments: three for the instance-level contrastive and temporal-classification objectives (96 frames) and one for multimodal pairs (32 frames, one frame per aligned utterance chunk). With the default $k = 8$, the chronological stream advances after eight total optimization steps; intermediate steps resample from replay. Visual and multimodal replay both retain a short FIFO sized to 10% of the stream and a long reservoir sized to 50%; multimodal minibatches draw one quarter of chunks from the short tier and the rest from the long tier. The instance, temporal-classification, and cross-modal losses each receive a coefficient of 0.5; the two visual losses therefore have twice the aggregate weight of the cross-modal loss.

CVCL baselines (offline vs. streaming replay schedules). All variants share the same encoder stack and a cross-modal InfoNCE objective on paired frames and utterances. Offline CVCL trains with shuffled minibatches for multiple epochs (batch size 8). *CL-CVCL* is BabyCL with $\mathcal{L}_I$ and $\mathcal{L}_S$ removed: it uses the identical chronological stream, online segmentation, dual replay buffers, batch composition, optimizer, learning-rate schedule, input normalization, and step count, differing only in that the visual losses are zeroed. The $k = 1$ and $k = 8$ variants differ only in the number of updates per buffer insertion. The stream is divided into nominal 3-minute windows (900 frames at 5 fps); after the first 406 windows (the short-tier capacity, about 10% of the stream) are pushed into the visual buffer in chronological order as a buffer warm-up with no parameter updates, the remaining $3{,}321$ windows are inserted one at a time with $k = 8$ updates each, giving $3{,}321 \times 8 = 26{,}568$ steps. BabyCL trains on 32 utterance–frame pairs per step (alongside 96 visual-buffer frames), so it sees $26{,}568 \times 32 \approx 850$k labeled-pair exposures, i.e. about 25 passes over the $33{,}737$ training pairs. The compute-matched offline CVCL therefore trains for 25 epochs i.i.d. (batch size 8, ≈ 105k steps); it matches BabyCL's labeled-pair exposure but takes about four times as many gradient steps. CL-CVCL runs the identical schedule ($26{,}568$ steps at $k = 8$; $3{,}321$ at $k = 1$) with the same 32 labeled pairs per step, so at $k = 8$ its labeled-pair exposure and gradient-step count both equal BabyCL's. Neither quantity is a guarantee on how often any individual example is revisited.

CVCL optimization. Offline CVCL uses AdamW (10^{-4} learning rate, weight decay 0.1) with ReduceLROnPlateau driven by the multimodal InfoNCE loss on a held-out validation split of $1{,}874$ utterances disjoint from the $33{,}737$ training utterances. CL-CVCL uses BabyCL's optimizer, learning-rate tiers, and warmup–cosine schedule (see above) and optimizes $\mathcal{L}_C$ only.

B Linear-probe details

Embeddings are ℓ_2-normalized and passed to a multinomial logistic regression with mild ℓ_2 regularization. The Labeled-S probe uses an 80/20 stratified train/test split over 26 categories; mini-ImageNet is fit on the canonical training split and scored on the validation split. Input normalization follows each model family (CLIP statistics for the CVCL family, ImageNet statistics for BabyCL). The Labeled-S and mini-ImageNet probes use the 512-d projected embedding; the BabyView probe (Appendix G) uses the pooled 2048-d backbone output.

C VTWT and Baby Winoground evaluation details

For both benchmarks, images follow the same preprocessing pipeline as the linear-probe evaluations. Text spans are embedded by averaging word vectors before projection to 512 dimensions; image and text embeddings are ℓ_2-normalized and compared by cosine similarity. A VTWT item is correct when the target phrase scores higher than the foil. For Baby Winoground, pairwise similarities define two ranking constraints, one under the original scene and one under the edited scene; Overall counts an item as correct only when both are satisfied.

D Backbone ablation details

We repeat the default BabyCL run with different vision backbones, keeping the rest of the training recipe fixed. All backbones are trained from scratch. Each architecture uses its own backbone learning rate and weight decay (Table 9); all other modules keep the default learning rate of 10^{-4}. CNN backbones follow the optimizer schedule of the main runs; ViT and ConvNeXt backbones use no weight decay on LayerNorm parameters and biases, a longer linear warmup (8% of steps) with a $0.1\times$ cosine floor, and gradient-norm clipping at 1.0. Each configuration is run with five random seeds.

Table 11: **Text encoder ablation.** Labeled-S 4AFC accuracy (single seed, end of training) when the default mean-pooled word-embedding text tower is replaced with a sequence encoder of the same output dimension.

Text encoder	4AFC (%)
BoW (default)	**43.76**
GRU	41.16
BiGRU	38.36
Tiny Transformer	32.99

Table 12: **Frozen CLIP text teacher.** Labeled-S 4AFC accuracy when the internal from-scratch text tower is replaced by a frozen CLIP ViT-B/16 text encoder, either as a post-hoc linear probe on a frozen BabyCL vision encoder or as the text tower during from-scratch training.

Training setup	Vision encoder	Text encoder	4AFC (%)
BabyCL ($k = 8$)	From scratch	From scratch	43.38 ± 1.47
Linear probe w/ frozen teacher	Frozen BabyCL (head only)	Frozen CLIP	$\mathbf{50.85 \pm 0.18}$
Frozen-teacher training	From scratch	Frozen CLIP	50.06 ± 0.15

E TEXT ENCODER ABLATION

The default text tower embeds an utterance as the mean of its learned word embeddings (BoW). We compare it against three sequence encoders trained from scratch over the same vocabulary, each producing the same 512-dimensional utterance embedding: a single-layer GRU, a single-layer bidirectional GRU, and a two-layer Transformer encoder. All other components of the default recipe are unchanged; each variant is run with a single seed, and we report end-of-training Labeled-S 4AFC accuracy.

As Table 11 shows, the mean-pooled BoW tower performs best and the more expressive sequence encoders perform progressively worse. The 4AFC probe queries single category words, which BoW embeddings—averages of word embeddings—match naturally; the sequence encoders instead learn whole-sentence representations, and during training they reach lower cross-modal InfoNCE loss while scoring worse on the single-word probe.

F TRAINING THE VISION ENCODER AGAINST A FROZEN TEXT TEACHER

BabyCL's text tower is a from-scratch word-embedding table with mean pooling. To estimate how much this weak text tower limits 4AFC performance, we replace it with a frozen CLIP ViT-B/16 text encoder (Radford et al., 2021) in two setups. In the *linear probe*, both the trained BabyCL vision encoder and the CLIP text tower are frozen, and we train only a linear head mapping the 512-dimensional visual embedding into CLIP's text space, with symmetric InfoNCE (temperature 0.07) on the training utterance–frame pairs (50 epochs, batch size 256, learning rate 10^{-3}, weight decay 10^{-4}); we report mean $\pm$ CI over five seeds of the head training on a single frozen checkpoint. In *frozen-teacher training*, we rerun the default recipe from scratch with the CLIP text encoder substituted for the internal text tower in $\mathcal{L}_C$; we report the mean $\pm$ CI of end-of-training checkpoints over five seeds. In both setups, 4AFC category names are embedded by the frozen CLIP text encoder.

As Table 12 shows, a linear head over the frozen BabyCL visual features reaches 50.85%, and from-scratch training against the frozen teacher reaches 50.06%, compared to 43.38% for the fully from-scratch model. The internal text tower, rather than the visual representation, is thus the main bottleneck on 4AFC. Since the frozen teacher brings in large-scale pretrained knowledge, these numbers serve as reference points rather than developmentally comparable results.

G BABYVIEW BENCHMARK DETAILS

We construct the BabyView 4AFC benchmark from 5,921 public, rater-validated object crops spanning 72 Communicative Development Inventory (CDI) noun categories. To match the vocabulary available to the SAYCam-trained text encoders, we retain the 61 categories present in the SAYCam vocabulary without synonym remapping. Using a fixed random seed, we generate 4,150 trials, each containing one target crop and crops from three distinct foil categories, following the Labeled-S

Table 13: **Credibility weighting.** End-of-training Labeled-S 4AFC (mean ± approximate 95% CI over five seeds) for the default recipe with and without credibility-weighted cross-modal pairs.

Training corpus	Default	+ credibility weighting
SAYCam	43.38 ± 1.47	**44.91 ± 1.39**
BabyView	28.54 ± 1.79	**32.93 ± 1.51**

image-mode scoring protocol. The companion linear-probe benchmark uses all 72 categories and a fixed stratified 80/20 split (4,734 training and 1,187 test crops).

Unlike the pre-squared Labeled-S evaluation images, the BabyView object crops vary substantially in aspect ratio and framing. We use shortest-side resize followed by a 224×224 center crop; channel normalization follows each model family's training statistics (CLIP statistics for the CVCL family and ImageNet statistics for BabyCL). Thus, the benchmark preserves the Labeled-S decision structure while introducing a separate corpus and more variable crop geometry, making it a challenging test of transfer rather than a difficulty-matched comparison.

H Credibility-weighted utterance–frame pairs

Child-directed speech often does not refer to what is visible, so many utterance–frame pairs carry no usable cross-modal signal; this is more severe for BabyView, whose transcripts are automatic. We therefore estimate a per-pair credibility offline with a pretrained CLIP ViT-B/32 and downweight low-credibility pairs during training.

Each aligned item i consists of an utterance u_i and its timestamp-defined video chunk. In sampling round r, let $\bar{s}_i^{(r)}$ be the mean CLIP cosine similarity between u_i and four frames sampled from its own chunk, and let $s_{ij}^{(r)}$ be the similarity between u_i and one frame sampled from the j-th of $K = 32$ other chunks. We define credibility as the fraction of foreign-chunk frames that the item's own frames outrank, averaged over $R = 3$ independent sampling rounds:

$$q_i = \frac{1}{RK} \sum_{r=1}^{R} \sum_{j=1}^{K} \mathbf{1}\left[\bar{s}_i^{(r)} > s_{ij}^{(r)}\right]. \tag{2}$$

Here, $\mathbf{1}[\cdot]$ equals one when its condition is true and zero otherwise, so $q_i \in [0, 1]$. Using win rates rather than raw similarity margins makes the score insensitive to CLIP's similarity scale. We map the pairwise chance level, $q_i = 0.5$, to zero weight and linearly increase the weight to one at $q_i = 0.9$:

$$w_i = \min\left(1, \max\left(0, \tfrac{q_i - 0.5}{0.4}\right)\right), \tag{3}$$

where $w_i \in [0, 1]$ is the credibility weight. Pairs at or below chance ($q_i \leq 0.5$) contribute zero weight to their own loss term, while pairs with $q_i \geq 0.9$ receive unit weight. All pairs remain in the minibatch and therefore still appear in the InfoNCE denominators as negatives. Let ℓ_i denote the average of the image-to-text and text-to-image cross-entropy contributions for pair i. The cross-modal loss over the current multimodal minibatch is

$$\mathcal{L}_C^w = \frac{\sum_i w_i \, \ell_i}{\max(\sum_i w_i, \epsilon)}, \tag{4}$$

where all sums run over that minibatch and $\epsilon = 10^{-8}$ prevents division by zero. The credibility scores are computed once before training. We train the default recipe with and without the weighting on both corpora, five seeds each; the unweighted runs are the paired controls from the same code state.

As Table 13 shows, the weighting helps on both corpora, and the gain is larger where transcripts are noisier: $+1.53$ points on SAYCam's hand-transcribed speech and $+4.39$ points on BabyView's automatic transcripts.

I Compute resources

All BabyCL online runs reported here were executed on single NVIDIA L40S GPUs. Under the default training configuration (batch size 128, mixed precision, and the replay schedule described above), observed peak device memory was on the order of 20 GB.

J LICENSES FOR EXISTING ASSETS

The original papers for all the code, data, and models used in the main experiments of the paper are cited. The terms are properly respected. SAYCam-related video and derived annotations are accessed through Databrary under authorized-access and data-use agreements (not redistributed). The vision backbone implementation follows the publicly released silicon-menagerie utilities (MIT License); all trained backbones are initialized from scratch; the pretrained checkpoints we use are kept frozen: the ImageNet ResNeXt-50 weights distributed with torchvision (temporal segmentation) and OpenAI CLIP ViT-B/32 and ViT-B/16 (MIT License; credibility weighting and the frozen-teacher reference). Baby Winoground and VTWT evaluation code/data labels follow the licensing of the BabyVLM evaluation repository: the core pipeline structure inherits the MIT License from lm-evaluation-harness, while the added multimodal tasks/models are released under Apache License 2.0. Core software dependencies include PyTorch and torchvision (BSD 3-Clause). Common scientific libraries used by our code are NumPy, SciPy, pandas, and scikit-learn (BSD 3-Clause), Pillow (HPND), and Matplotlib (PSF-based).