Looking to Learn: Token-wise Dynamic Gating for Low-Resource Vision-Language Modelling

Bianca-Mihaela Gănescu Suchir Salhan Andrew Caines Paula Buttery

ALTA Institute & Computer Laboratory, University of Cambridge

TL;DR

Problem: Multimodal models trained under BabyLM constraints struggle to integrate limited visual information effectively.

Solution: A lightweight decoder-based model with a **token-wise dynamic gate** that adaptively fuses linguistic and visual cues.

Findings: Our model achieves competitive or superior performance on five BabyLM benchmarks and learns *interpretable gating patterns*, favouring vision for generally more grounded words and text for function words.

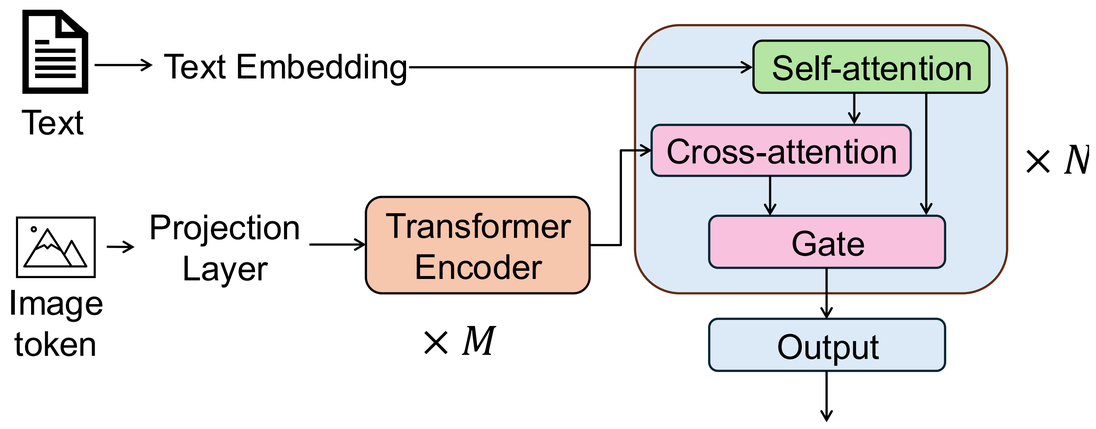
Methodology

1. We train a lightweight dual-stream decoder that fuses text and vision through a **token-wise dynamic gate**:

$$h_{\text{fused}} = g \odot h_{\text{text}} + (1 - g) \odot h_{\text{crossAttn}}$$

2. To maximise limited visual information (*CLS* token), we investigate **feature modulation techniques** & channel attention.
3. For visual grounding, we test two auxiliary objective contrastive learning functions, **CLIP** & **LexiContrastive (LCG)**, alongside next-token prediction.

Architecture



Key Results

Model	BLiMP	BLiMP-S	EWoK	Winoground	VQA*
BabyLM Challenge 2025 Baselines					
Flamingo	70.9	65.1	51	54.8	43.31
GIT	72.2	66.4	51.8	56.2	49.82
Our framework					
Base, soft gate per feature	74.33	56.36	50.81	51.61	50.02
Architectural features					
Soft gate per token	73.86	55.43	51.56	52.14	48.39
Hard gate per feature	74.10	54.16	51.20	50.13	45.62
Hard gate per token	74.19	54.59	51.16	50.80	45.51
No gate	74.70	55.75	50.77	51.34	50.58
FiLM on text	74.32	55.10	50.61	53.49	46.04
FiLM on cross-attention	74.95	56.36	51.62	52.68	49.66
FiLM on image	73.80	54.59	51.06	50.13	17.92
DyIntra on text	74	56.97	51.73	51.47	47.16
DyIntra on cross-attention	73.68	56.68	51.57	53.22	48.87
DyIntra on image	74.69	56.57	51.28	50.00	45.61
Channel attention	74.24	54.23	51.15	51.15	49.15
Auxiliary objective functions					
NTP + CLIP	72.28	54.35	51.45	51.47	47.72
NTP + LCG	70.27	56.91	49.74	50.00	36.62

Table 1. **Performance of our base model and variants on five BabyLM benchmarks.** Scores are computed using the 2025 (BLiMP, BLiMP-S, EWoK, Winoground) and 2024 (VQA) evaluation pipelines. Green shading indicates performance above the baselines.

Result #1: Our framework achieves a higher score on BLiMP (~4% higher than Flamingo and >2% higher than GIT) and competitive scores for EWoK, Winoground and VQA.

Result #2: The dynamic gating modules match our *no gate* model's performance on BLiMP and BLiMP-S, modestly improve on Winoground and show mixed results on VQA.

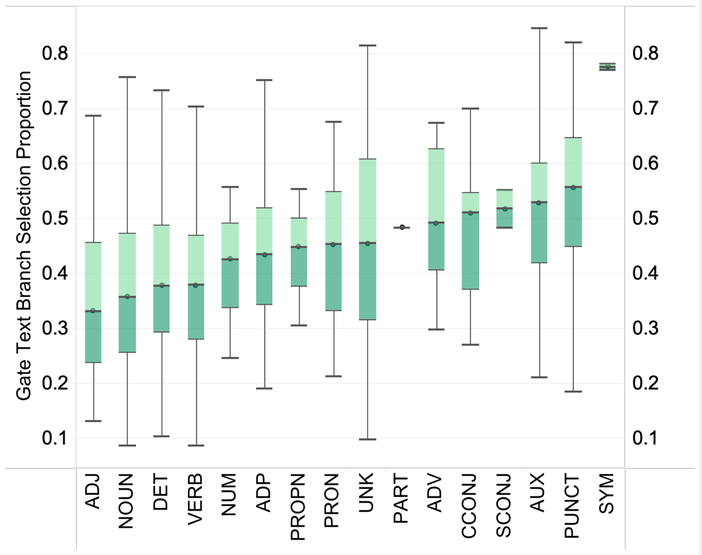
Results #3: Modulation and channel attention achieve mixed results over the five benchmarks, underscoring that the global image embedding represents a performance bottleneck.

Results #4: A pure next token prediction objective function achieves the best scores for our base model overall.

Paper

Code

Interpretability



Gate Value per Concreteness Bin	
Category	**Mean (SD)**
Very Abstract (<318)	0.427 (0.141)
Abstract (318–438)	0.471 (0.136)
Concrete (438–558)	0.391 (0.155)
Very Concrete (>558)	0.343 (0.139)

Gate Value per Imageability Bin	
Category	**Mean (SD)**
Very Low (<342)	0.427 (0.143)
Low (342-450)	0.481 (0.130)
High (450-558)	0.378 (0.126)
Very High (>558)	0.351 (0.155)

For the PoS which are open-class and generally more grounded (ADJ, NOUN, VERB) (Haley et al., 2025), the model attends more to the image signals (left side of the plot), while for function words (CONJ, AUX, PART) the model attends more to the pure text.

Conclusions and Future Work

We show that token-wise dynamic gating enables small vision-language models to adaptively integrate linguistic and visual cues, yielding interpretable patterns and competitive performance.

We identify four key directions for **future work**:

More diverse and high-quality data for training, covering constructions and concepts present in the evaluation benchmarks.

Fully multimodal training dataset for training stability and improved language acquisition.

Patch-token image embeddings for richer multimodal learning.

Rewarding cognitive-plausible mechanisms i.e., evaluating the cognitive principles guiding the models responses.