Towards understanding multimodal in-context learning

Yiran Huang[1,2,3] **Karsten Roth**[2,3,4] **Quentin Bouniot**[1,2,3] **Wenjia Xu**[5] **Zeynep Akata**[1,2,3]
[1]Technical University of Munich, Germany
[2]Helmholtz Munich, Germany
[3]Munich Center for Machine Learning (MCML), Germany
[4]Tübingen AI Center, University of Tübingen, Germany
[5]Beijing University of Posts and Telecommunications, China

Abstract

Multimodal Large Language Models (MLLMs) exhibit in-context learning (ICL) abilities. Yet we lack understanding of how these models actually perform multimodal ICL. We train modern transformer models on synthetic classification tasks, systematically varying data statistics and model architecture. We find that pretraining on a highly diverse primary modality installs the ICL circuit, so the secondary modality can attain comparable ICL with much less data complexity. Scaling up multimodal decoders improves ICL capacity while the encoder for the second modality sets the ceiling, as weak representations bottleneck multimodal ICL. Rotary position embeddings (RoPE) actively harm ICL by disrupting attention circuits with fixed data complexity. Through mechanistic analysis with progress measurements that track the formation of ICL circuits, we demonstrate that both unimodal and multimodal ICLs rely on a common induction-style circuit that copies the label from the in-context exemplar that matches the query. Multimodal training primarily refines this behavior rather than introducing new circuitry. These results offer a clear, mechanism-level account and practical levers for engineering ICL in modern multimodal transformers.

1 Introduction

In-Context Learning (ICL) has emerged as a remarkable capability of Large Language Models (LLMs)[2], with MLLMs extending this ability across modalities [1]. While prior work has examined ICL in simplified unimodal transformer settings [3, 8, 11], the driving factors and mechanisms underlying multimodal ICL remain poorly understood. In this work, we address this gap through simple, controlled in-context classification tasks and study: (1) transfer of data distributional drivers and architectural factors for ICL from unimodal to multimodal regimes, (2) how the second-modality representation impacts cross-modal ICL, and (3) transferability of circuit-level mechanistic interpretations (previous-token & induction heads) via progress measurements

2 Methodology

We investigate ICL through controlled experiments using synthetic Gaussian Mixture Models (GMMs) for both modalities, allowing precise control over data properties including class diversity K, burstiness B, within-class noise ϵ, and Zipfian distribution parameters α (detailed in Suppl. A.1). In the unimodal setting, models receive N item-label pairs followed by a query. For multimodal tasks, we provide paired exemplars from two modalities: $x_1, x'_1, \ell_1, x_2, x'_2, \ell_2, \ldots, x_N, x'_N, \ell_N, x_q, x'q$. For evaluation, we distinguish between In-Weight Learning (IWL) using memorized knowledge and

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

ICL requiring contextual reasoning by testing on novel classes or permuted labels (Suppl. A.1). We employ a 2-layer transformer with RMSNorm [17], SiLU activation [5], and optional RoPE [13], more closely aligned with modern LLMs than previous work [11]. For multimodal tasks, we append an MLP projector on the pretrained decoder. For encoder studies, a pretrained modality-2 encoder is inserted before the projector (Suppl. A.2).

3 Data distributional and model structural findings in multimodal ICL

Data statistics still govern multimodal ICL. Prior work [11] showing that high burstiness, class diversity, and within-class noise maximize unimodal ICL, we find these principles transfer to multimodal settings with an asymmetry. Initializing decoders from unimodal pretraining on highly diverse primary modality data, strong multimodal ICL emerges even when secondary modality data has only moderate complexity (Suppl. A.3). This suggests the primary modality installs core ICL circuits that secondary modalities can leverage with significantly less training complexity.

Scaling up improves multimodal ICL Scaling the decoder consistently improves multimodal ICL by increasing the representational bandwidth needed to integrate secondary modalities. (Suppl. A.4.1).

Rope impairs multimodal ICL. RoPE consistently impairs ICL compared to absolute encodings given fixed data complexity because its multiplicative nature obscures the discrete positional cues required for induction heads and token copying.(Suppl. A.4.2).

4 The encoder plays an important role in multimodal ICL

Encoder reshapes modality two. A pretrained encoder significantly boosts ICL by compressing secondary modality features into a low-rank subspace and aligning them with the primary modality, effectively removing representational bottlenecks for induction (Suppl A.6.2).

Better perception leads to better ICL also in real images. Experiments on Omniglot show that higher encoder accuracy correlates with improved ICL. However, fully exploiting strong encoders requires joint optimization (updating the decoder), as training only the encoder or projector quickly plateaus(Suppl. A.6.3).

5 Progress measurements for multimodal ICL

We quantify the formation of ICL circuits with two minimal, layer-specific metrics: *(i)* $\mathrm{PHStrength_m}$ measures the mean attention paid by the tokens to their immediate predecessors in layer m, and *(ii)* $\mathrm{IndStrength_m}$ quantifies how strongly the target token attends to labels directly preceded by the context examples from the same class in layer m (Suppl. A.5). Across multimodal models, Pearson correlations [9] with ICL accuracy are strong for $\mathrm{IndStrength_2}$ ($r \approx 0.70$) and substantial for $\mathrm{PHStrength_1}$ ($r \approx 0.58$). Using only $\mathrm{PHStrength_1}$ and $\mathrm{IndStrength_2}$ also yields high predictive power for ICL ($R^2 \approx 0.90$) using Random Forest regressor, establishing these measurements as compact, mechanism-aligned diagnostics for multimodal ICL. The findings reveal the dominant driver of continued ICL gains is the induction head in the second layer. The previous-token head, while still a strong signal, no longer tracks ICL improvement as closely because it has already stabilized from unimodal pretraining. This suggests that multimodal ICL depends less on learning a new ICL task and more on refining the label retrieval and matching process across modalities.

6 Conclusion

This work provides definitive answers to three fundamental questions about multimodal ICL: statistical drivers transfer robustly but with critical asymmetry—the primary modality needs high complexity while the secondary achieves comparable ICL with much less class diversity; RoPE consistently impairs ICL by disrupting attention circuits while scaling improves ICL; the encoder for the second modality sets the ceiling, as weak representations bottleneck multimodal ICL. Our interpretable progress measurements reliably predict ICL performance and establish practical diagnostics for model development.

References

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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in neural information processing systems*, 35:18878–18891, 2022.

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

[5] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *arXiv preprint arXiv:1702.03118*, 2018.

[6] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[7] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. The omniglot challenge: a 3-year progress report. *Current Opinion in Behavioral Sciences*, 29:97–104, 2019.

[8] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *CoRR*, 2022.

[9] Karl Pearson and Francis Galton. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895. doi: 10.1098/rspl.1895.0041. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspl.1895.0041`.

[10] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

[11] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=aN4Jf6Cx69`.

[12] Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *International Conference on Machine Learning*, pages 45637–45662. PMLR, 2024.

[13] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Zhai. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2022.

[14] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[16] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No" zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[17] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in neural information processing systems*, volume 32, pages 1–11, 2019.

[18] Yongshuo Zong, Ondrej Bohdal, and Timothy M Hospedales. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *CoRR*, 2024.

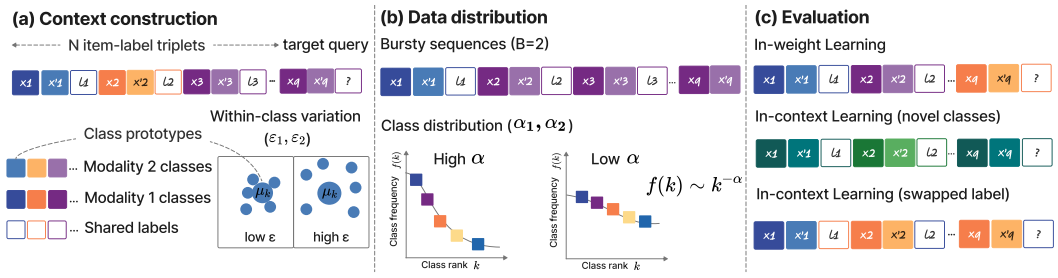

Figure 1: Overview of the preliminaries in the multimodal setting. **(a)** The context consists of N triplets followed by the target query. The paired examples (x_i, x_i') from two modalities, with a shared label l_i, are generated from GMMs by controlling within-class variation ε_1 and ε_2. **(b)** The distributional properties for the synthetic data from both modalities. The burstiness B determines repeated class occurrences in context. Both class frequencies follow a Zipfian distribution with exponent α_1 and α_2. **(c)** Evaluation distinguishes between IWL, where target queries belong to class seen during training while not in the context during evaluation, and ICL, where target queries are novel but in the context. A swapped-label condition further isolates ICL by permuting the labels.

A Technical Appendices and Supplementary Material

A.1 Preliminaries

A.1.1 Task description

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ denote the input spaces of two modalities and let $\mathcal{L} = \{1, \ldots, L_1\}$ be the shared label set. We provide the model with a context consisting of N labelled exemplars followed by an unlabelled query. Unless otherwise stated, N is fixed to 8 throughout the paper.

Unimodal setting. Following Reddy [11], the context comprises N item–label pairs from a single modality: $x_1, \ell_1, x_2, \ell_2, \ldots, x_N, \ell_N, x_q$. Each $x_i \in \mathcal{X}_1$ is an example whose ground-truth label is $\ell_i \in \mathcal{L}$. The model must predict ℓ_q for the query item x_q.

Multimodal setting. We extend the task by presenting paired exemplars from two modalities: $x_1, x_1', \ell_1, x_2, x_2', \ell_2, \ldots, x_N, x_N', \ell_N, x_q, x'q$ (Fig. 1a). Here $x_i \in \mathcal{X}_1$ and $x_i' \in \mathcal{X}_2$ correspond to the same label ℓ_i. Importantly, at least one exemplar (unimodal) or exemplar pair (multimodal) shares the query's class, ensuring that ICL is in principle possible.

A.1.2 Synthetic Data Generation

We generate data for both modalities from Gaussian Mixture Models (GMMs), allowing for precise control over data properties. For clarity, we describe the procedure for a single modality $m \in \{1, 2\}$ and omit the superscript m. Items are sampled from a GMM with K classes. Each class k is defined by a prototype vector $\mu_k \in \mathbb{R}^D$, with coordinates drawn independently from $\mathcal{N}(0, 1/D)$. An instance of class k_1 is obtained by adding Gaussian noise to its prototype:

$$x_i = \frac{\mu_k + \varepsilon\eta}{\sqrt{1+\varepsilon^2}}, \quad \text{where} \quad \eta \sim \mathcal{N}(0, I_D/D). \tag{1}$$

The parameter ε_1 sets the within-class variability (Fig. 1a). The rescaling factor ensures that $\|x\| \approx 1$. We then map the K classes to L labels, where $L_1 \leq K_1$. This simulates real-world scenarios where many distinct instances or subclasses map to a single semantic concept. Each label is also associated with a prototype vector sampled from $\mathcal{N} \sim (0, I_D/D)$. For the multimodal setting, we generate data for the two modalities using separate GMMs with potentially different parameters (K_1, ε_1 for modality 1; K_2, ε_2 for modality 2). Crucially, the label set for the second modality is a subset of the first ($L_2 \leq L_1$), mimicking typical multimodal models (e.g., MLLMs) where the output space is determined by one primary modality (i.e., text).

A.1.3 Parameterizing the data distribution

Beyond the intrinsic GMM parameters (D, K, ε, L), the data distribution for our training and test sequences is governed by the following hyperparameters, which were introduced by [11] to study

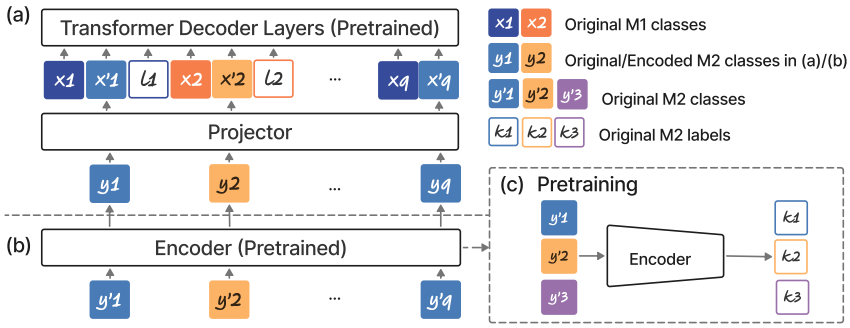

Figure 2: Multimodal setup extending the unimodal decoder. The decoder is initialized from the unimodal checkpoint. M1 denotes the base modality and M2 the additional modality. *(a) Projector-only setup*: an MLP projector aligns features to the M1 embedding space; projector and decoder are trained jointly. *(b) Encoder-augmented setup*: a pretrained M2 encoder is stacked before the projector and decoder. *(c) Encoder pretraining task*: the M2 encoder is pretrained on M2-specific classes/labels (not the shared multimodal labels).

ICL dynamics (see Fig. 1b). The burstiness B specifies the number of occurrences of items from a particular class in the input context. For such sequences, the number of examples N is a multiple of B, and the context consists of N/B classes, each appearing exactly B times. p_B determines how often such bursty sequences occur: with probability p_B an input is generated in the bursty manner as described, whereas with probability $1 - p_B$ the N context items and the query are sampled i.i.d. across classes. The rank-frequency distribution over the classes follows Zipfian's law [10], which is denoted as $f(k) \sim k^{-\alpha}$ with exponent $\alpha > 0$ controlling the skew of this distribution.

A.1.4 Evaluation Metrics

We assess model performance by disentangling two learning mechanisms: IWL, which relies on knowledge stored in model parameters, and ICL, the ability to reason from contextual examples at test time. We illustrate the evaluation in Fig. 1c. To assess IWL, we evaluate the model on test sequences where context and query items are sampled i.i.d. from the training distribution. Success on this task depends on the knowledge stored in the model's weights. To assess ICL, we generate sequences whose context and query classes are entirely novel, forcing the model to rely on the context. As an additional ICL metric we evaluate the model on sequences in which the context labels are permuted relative to training, invalidating the memorized mapping.

A.2 Methodology

A.2.1 Establishing Architectural Premises: ICL in Modern Transformers

While previous work has revealed foundational properties of ICL, such as dependency on specific data distribution [11, 3], mechanisms that implement ICL algorithms [8], many questions remain, particularly regarding how these principles operate within modern architectures. Firstly, we construct a controlled unimodal setting to isolate the data-centric and architecture-centric factors that drive ICL, providing essential premises for our later multimodal study. To achieve this, we introduce a two-layer transformer architecture incorporating RMSNorm [17], SiLU activation [5], and RoPE [13]. While this configuration aligns more closely with the architectural choices of contemporary LLMs [15], it diverges from architectures used in previous mechanistic analyses (e.g., Reddy [11]). This deliberate design choice enables us to examine whether previously identified data properties that foster ICL in simplified settings remain effective within more realistic, LLM-inspired configurations.

A.2.2 Multimodal experimental setup

We retain the same transformer decoder architecture and append a modality-specific MLP projector that maps features from the second modality into the embedding space of the first. We initialize the transformer with a pretrained checkpoint from the unimodal stage and then jointly train the projector

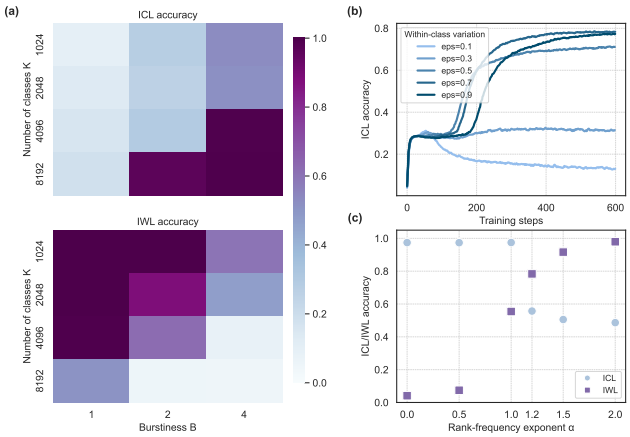

Figure 3: Data distributional results in the modern unimodal setting. (a) Increasing the number of classes K, burstiness B promotes ICL. (b) Increasing within-class variation ε promotes ICL. (c) Increasing the Zipf exponent α shifts performance toward IWL, with an optimal balance achieved when $\alpha = 1$.

and the decoder. We denote the modality used in unimodal pretraining as modality 1 (M1) and the additional modality as modality 2 (M2)(Fig. 2).

A.3 Data distributional findings

We begin by examining how the statistical properties of the training data affect the emergence of ICL in our modern two-layer transformer, and by comparing these behaviors to those reported in simplified architectures [11, 3]. All key trends are reproduced (Fig. 3): increasing the number of classes K, burstiness B or within-class variation ε promotes ICL, while increasing the Zipf exponent α shifts performance toward IWL, with an optimal balance achieved when $\alpha = 1$ (Fig. 3c). These findings confirm that the core statistical drivers are architecture-agnostic. Our experiments reveal two extensions beyond prior work. First, strong ICL does not require perfect burstiness: even with $p_B < 1$, high K and B suffice, indicating that induction behavior learned from structured contexts transfers to unstructured ones. Second, while higher within-class noise ε again promotes ICL, we additionally find that it slows convergence (Fig. 3b), as the model requires more data to form robust associations.

Extending to the multimodal setting, Consistent with the unimodal case, increasing burstiness (B) and the number of classes (K_1,K_2) improves ICL performance while decreasing IWL. In our experiments, we fix $K_1 = 8192$ and vary K_2 to probe the role of diversity in modality 2. Surprisingly, relatively small K_2 suffices: with $K_2 = 256$ and $B = 4$, the model already reaches near-perfect ICL accuracy (Fig. 4a). This result stems from the asymmetric roles of the two modalities. Since modality one has high diversity and is used to pretrain the decoder, it provides a strong foundation for ICL. Modality two, therefore, doesn't need high class granularity to induce ICL; its primary role is to provide a distinguishable signal that can be mapped into the pre-established embedding space of modality one. Both within-class noise parameters, ε_1 and ε_2, positively contribute to ICL, but increasing ε_2 has a substantially stronger effect (Fig. 4b). Modality one is already well-represented from unimodal pretraining, so additional variability adds limited new information. In contrast, higher ε_2 forces the model to learn a more robust, generalizable mapping from modality two to the shared embedding space, promoting ICL. For the class distribution, when the decoder is pretrained on modality 1 with Zipfian skew $\alpha_1 = 1$, we observe the best multimodal balance near $\alpha_2 = 1$ for modality 2 as well (Fig. 4c); this holds across a range of K and B. This result aligns with empirical observations from real datasets. Udandarao et al. [16] reveals that the concept distribution of pretraining datasets including CC-3M, CC-12M [4], and YFCC-15M [14] is highly long-tailed. Both vision-only and vision-language datasets exhibit highly skewed distributions, suggesting that skewed class distributions in the second modality provide an effective complement to the naturally skewed distributions found in large-scale pretraining data.

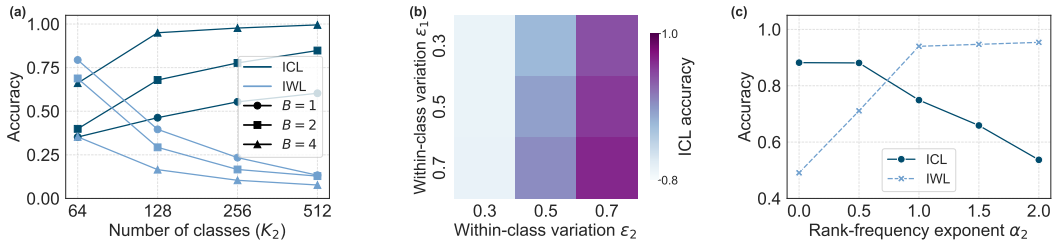

Figure 4: Effects of multimodal data distributions on ICL and IWL. *(a) Class diversity and burstiness.* With $K_1 = 8192$, ICL accuracy increases with K_2 and B; a surprisingly small class diversity in the second modality ($K_2 = 256, B = 4$) is sufficient for near-perfect ICL accuracy. *(b) Within-class variation.* Heatmap of ICL over $(\varepsilon_1, \varepsilon_2)$ shows increasing ε_2 has a markedly stronger positive effect than increasing ε_1. *(c) Class-distribution skew.* With α_1=1 fixed, the ICL–IWL balance is best when $\alpha_2 \approx 1$, i.e., when modality 2's skew matches modality 1's long-tailed distribution.

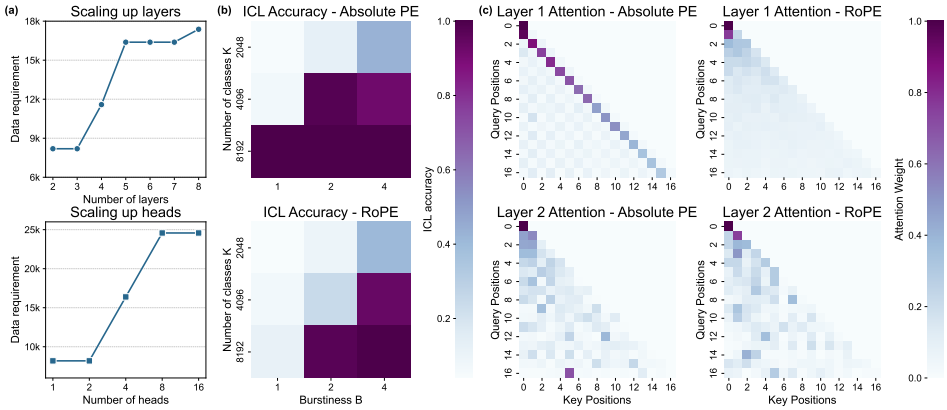

Figure 5: The impact of model architecture on unimodal ICL. *(a) Data requirement (measured by $K \cdot \sqrt{B}$) needed for larger modes to recover ICL.* Scaling up the model in depth and width both are shown to hurt ICL. *(b) ICL accuracy for models with different PE.* RoPE decreases ICL accuracy compared with absolute PE across different combination of data complexity. *(c) The attention pattern in the two-layer model with absolute PE and RoPE.* The correct label is in position 5. With absolute PE, both the previous token head and the induction head pattern are clearly observable while they are less visible with RoPE.

A.4 Model architectural findings

A.4.1 Model scaling impacts ICL

We systematically scale model depth and attention heads while keeping training data constant, revealing a fundamental tension between model capacity and ICL emergence. Larger models consistently exhibit reduced ICL performance, requiring increasingly complex training data (measured by $K \cdot \sqrt{B}$) to recover lost capabilities(Fig. 5a). This requirement grows faster with the number of heads than with the number of layers. We attribute this to how multi-head attention distributes capacity: more heads allow information to be partitioned into specialized subspaces, enabling the item–label memorization with little pressure to coordinate previous-token + induction circuit. This creates a low-loss shortcut that competes with the algorithmic ICL solution. Adding layers increases depth but does not create as many independent storage slots, so the bias toward in-weight learning is weaker. This interpretation aligns with prior observations that induction-style behavior typically emerges in a subset of heads while other heads specialize in memorization [6, 12]. Importantly, scaling does not *eliminate* ICL capacity; it raises the threshold of structured signal needed for the algorithmic solution to outcompete memorization, which we achieve by increasing K and B in the training distribution.

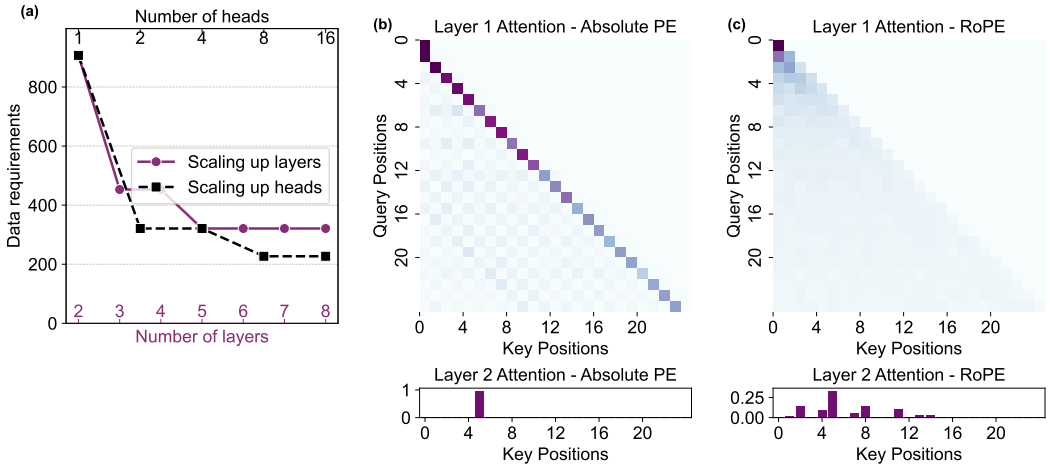

Figure 6: Impact of model architecture on ICL. (a) Data requirements measured by needed for larger multimodal model to achieve perfect ICL. Larger decoders (more layers or attention heads) achieve the same accuracy with lower data requirements. (b) Attention comparison of absolute PE in multimodal setting. (c) Attention comparison of RoPE in multimodal setting. Correct label is in position 5. The previous token head and induction head are clearly visible with absolute PE while not with RoPE.

Unlike the unimodal case, where scaling without stronger data cues biased models toward IWL, larger decoders require less complex data to achieve ICL (Fig. 6a) in the multimodal setting. Because the decoder already performs ICL over modality one, scaling primarily increases representational bandwidth for integrating modality two via the projector, rather than relearning ICL from scratch. Larger models better accommodate cross-modal variability and generalize more effectively even with fewer classes or less burstiness in the training data.

A.4.2 RoPE hurts ICL

Switching from absolute positional encodings (PE) to RoPE causes a marked drop in ICL accuracy (Fig. 5b). Attention visualizations confirm that the model struggles to form strong previous-token heads with RoPE and the induction head is also less clear (Fig. 5c). We hypothesize this is because RoPE's multiplicative rotational structure lacks the discrete, offset-based cues of absolute encodings, making it harder for the model to learn the simple token-copying operations crucial for ICL. While RoPE improves scalability, it may suppress inductive biases needed for ICL.

While our unimodal findings showed that ICL relies on a circuit of previous-token and induction heads, a key question is whether these same mechanisms are responsible for multimodal ICL. We first observe that models with absolute PE successfully form these circuits (6b), indicating that the core ICL mechanism learned during unimodal pretraining successfully transfers to the more complex multimodal setting. In contrast, RoPE significantly reduces ICL accuracy in this setting (Appendix). Attention visualizations (6c) confirm that RoPE hinders the formation of key ICL mechanisms, which is consistent with unimodal setting. The fundamental reason remains the same: RoPE's relative encoding interferes with the clear, offset-based cues that are essential for the formation of the previous-token and induction heads.

A.5 Progress measurements

The ICL mechanism in small transformers has been traced to a circuit of specialized attention heads: a previous-token head copies information from the previous token, and an induction head uses this to locate matching examples and find tokens preceded by the matching tokens [8]. Building on this, we design a suite of "progress measurements" to quantify the formation of these circuits, especially when attention patterns are vague, as with RoPE Except for Context Label Accuracy, all metrics are computed separately for each transformer layer $m \in \{1, 2\}$. The attention at layer m is denoted by $\mathrm{Attn}_{\mathrm{m}}$.

Previous Token Head Strength ($\mathrm{PHStrength}$) measures attention paid by all the tokens to their immediate predecessor. Given the sequence length L_{seq}, we define:

$$\mathrm{PHStrength}_{\mathrm{m}} = \frac{1}{L_{\mathrm{seq}} - 1} \sum_{i=1}^{L_{seq}-1} \mathrm{Attn}_{\mathrm{m}}\big(\text{query}_i \to \text{key}_{i-1}\big). \tag{2}$$

Induction Head Strength ($\mathrm{IndStrength}$) quantifies how strongly the target token attends to tokens directly preceded by the context examples from the same class.

$$\mathrm{IndStrength}_{\mathrm{m}} = \frac{1}{|\mathcal{P}|} \sum_{j \in \mathcal{P}} \mathrm{Attn}_{\mathrm{m}}(\text{query}_t \to \text{key}_j), \tag{3}$$

where $\mathcal{P}$ is the set of positions immediately after context examples of the same class as the target.

Target Label Association (TLA) measures total target attention to all context label positions.

$$\mathrm{TLA}_{\mathrm{m}} = \sum_{k \in \mathcal{Y}} \mathrm{Attn}_{\mathrm{m}} \left(\text{query}_{\mathrm{t}} \to \text{key}_j\right), \tag{4}$$

where $\mathcal{Y}$ is the set of positions of all the labels in the context.

Context-label accuracy (CLA) measures if the predicted label appeared in the context, indicating reliance on context.

$$\mathrm{CLA} = \mathbb{P}\left(\hat{y} \in \{y_i\}_{i=1}^{N}\right), \tag{5}$$

where $\hat{y}$ is the predicted label, $\{y_i\}_{i=0}^{N}$are the labels are the labels of the N context examples.

In the multimodal setting, inputs are provided as triplets, so we define $\mathrm{PHStrength}^{(1)}$ as the attention a token pays to its immediate predecessor and $\mathrm{PHStrength}^{(2)}$ for attention to the token two positions earlier. This allows us to investigate if the model shifts its attention to wider offsets. Table 2 shows the Pearson correlation between our metrics and ICL accuracy.

A.5.1 Pearson Correlation

To validate our progress measurements, we compute the Pearson correlation between each metric and ICL accuracy across a large pool of trained models. The results (Table 1) align with mechanistic interpretations. PHStrength in layer 1 shows the strongest correlation, confirming that early-layer previous-token heads are critical. IndStrength in layer 2 is also strongly correlated, consistent with induction head performing the final label-retrieval step. CLA ranks second, showing that ICL-capable models learn to predict labels from the context. A moderate correlation for TLA in layer 1 suggests it aids a general label-selection mechanism.

Table 1: Pearson correlations (r) between progress measurements (PM) and ICL accuracy in the unimodal setting, ranked by $|r|$. We consider $|r| > 0.5$ strong and $|r| > 0.7$ very strong.

Rank	PM	r
1	$\mathrm{PHStrength}_1$	0.72
2	CLA	0.65
3	$\mathrm{IndStrength}_2$	0.61
4	TLA_1	0.59
5	TLA_2	0.11
6	$\mathrm{IndStrength}_1$	0.06
7	$\mathrm{PHStrength}_2$	-0.10

In the multimodal setting, inputs are provided as triplets, so we define $\mathrm{PHStrength}^{(1)}$ as the attention a token pays to its immediate predecessor and $\mathrm{PHStrength}^{(2)}$ for attention to the token two positions earlier. This allows us to investigate if the model shifts its attention to wider offsets. Table 2 shows the Pearson correlation between our metrics and ICL accuracy.

We find that $\mathrm{PHStrength}_1^{(1)}$ remains a strong signal (Pearson $r = 0.57$), while $\mathrm{PHStrength}^{(2)}$ shows consistently low correlation across both layers. This suggests that even in multimodal input sequences, the model does not fundamentally shift its attention span, and still relies primarily on standard previous-token copying mechanisms developed during unimodal pretraining. $\mathrm{IndStrength}_2$ shows the highest correlation, indicating that improvements in ICL during multimodal training are most strongly associated with refinement in the model's ability to attend to the correct label token. During the unimodal pretraining stage, the model already learns a strong previous-token head in the first layer which is retained in the multimodal stage, so its correlation with new ICL gains is weaker. The dominant learning objective in the multimodal stage is to match the new modality to the correct class, making the continued development of the induction head the most sensitive indicator of progress.

Table 2: Pearson correlation (r) between progress measurements and ICL accuracy in the multimodal setting.

Rank	Metric	r
1	$\mathrm{IndStrength}_2$	0.70
2	$\mathrm{PHStrength}_1{}^{(1)}$	0.58
3	TLA2	0.56
4	TLA_1	0.51
5	$\mathrm{PHStrength}_1{}^{(2)}$	0.48
6	$\mathrm{PHStrength}_2{}^{(2)}$	0.47
7	$\mathrm{IndStrength}_1$	0.10
8	CLA	0.02
9	$\mathrm{PHStrength}_1{}^{(1)}$	-0.02

A.5.2 ICL predictable from progress measurements

To assess sufficiency in addition to correlation, we train a random-forest regressor to predict ICL accuracy from the progress measurements. As shown in Table 3, the results are compelling. The two core mechanistic components, $\mathrm{PHStrength}_1$ and $\mathrm{IndStrength}_2$ are sufficient to predict ICL accuracy with high fidelity ($R^2 = 0.909$). This indicates they capture the primary mechanism for ICL. Adding TLA_1 further boosts performance ($R^2 = 0.960$), confirming its role as a complementary mechanism for locating context-relevant labels.

Table 3: Random-forest performance (R^2) for predicting unimodal ICL accuracy from subsets of progress measurements.

Feature subset	$\mathbf{R^2}$ (mean $\pm$ std)
$\mathrm{PHStrength}_1,\ \mathrm{IndStrength}_2$	0.909 ± 0.023
$\mathrm{PHStrength}_1,\ \mathrm{TLA}_1,\ \mathrm{IndStrength}_2$	0.960 ± 0.022
All features	0.974 ± 0.014

Similarly, we train a random forest regressor to predict ICL accuracy and the results are shown in the Table 4. We first confirm that the core induction mechanism remains the primary driver: using just $\mathrm{PHStrength}_1^{(1)}$ and $\mathrm{IndStrength}_2$ achieves a strong predictive performance with $R^2 = 0.9$. Adding TLA_1 improves the R^2 to 0.96. In contrast, adding TLA_2 results in a smaller gain ($R^2 = 0.92$), which supports our interpretation that its contribution is largely redundant with $\mathrm{IndStrength}_2$. Using all available metrics yields $R^2 = 0.98$, showing that multimodal ICL accuracy can be predicted with high fidelity from a small set of interpretable attention behaviors.

Table 4: RF performance (R^2) using subsets of the progress measurements in the multimodal setting.

Feature subset	$\mathbf{R^2}$ (mean $\pm$ std)
$\mathrm{PHStrength}_1^{(1)}$, $\mathrm{IndStrength}_2$	0.90 ± 0.01
$\mathrm{PHStrength}_1^{(1)}$, $\mathrm{IndStrength}_2$, TLA_1	0.96 ± 0.06
$\mathrm{PHStrength}_1^{(1)}$, $\mathrm{IndStrength}_2$, TLA_2	0.92 ± 0.01
All metrics	0.98 ± 0.01

A.6 The Role of the Encoder and Its Impact on Feature Representations

Our unimodal and multimodal analyses have demonstrated that the transformer's ability to form specific attention heads is critical for ICL. However, these mechanisms can only operate on the representations they receive. A crucial question is: how does the quality of the input representation itself affect a multimodal model's ability to perform ICL?

A.6.1 Pretrained Encoders Enhance Multimodal ICL

We investigated the above question by manipulating the feature representation of the modality-two inputs. In initial experiments, we observed a consistent drop in ICL performance as the feature dimensionality of modality-two data increased (Figure 7a). This suggested that the model struggles to extract task-relevant structure from high-dimensional inputs when no dedicated encoder is available. To address this, we introduce an encoder pretrained on modality two data. We connect this encoder to the decoder via an MLP projector. We experiment with three training strategies: (1) training only the projector, (2) training the projector & pretrained decoder, and (3) training all the components.

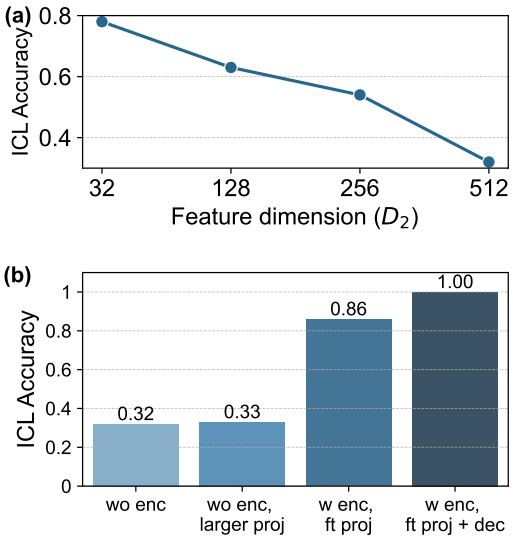

Figure 7: Pretrained encoders enhance multimodal ICL. *(a) Sensitivity to feature dimension:* ICL accuracy decreases as the second modality's feature dimension increases. *(b) Encoder vs. capacity:* Encoder-equipped models consistently outperform capacity-matched projector-only models.

Results shown in Figure 7b indicate that introducing a pretrained encoder significantly improves ICL accuracy across all training regimes. This supports the hypothesis that structured representations from modality-two are critical for enabling effective cross-modal matching and label induction. To ensure that the performance gain from adding a modality-two encoder is not merely due to increased parameter count, we conduct a controlled comparison. Specifically, we scale up the MLP projector so that the total number of parameters in the encoder-free model matches that of the encoder-equipped model. This setup allows us to isolate the effect of structural inductive bias introduced by the encoder from that of model capacity alone. As shown in Figure 7b, the model with a pretrained encoder consistently outperforms the one with an enlarged projector, despite having a similar parameter budget. This result confirms that the encoder contributes meaningfully to representation quality and cross-modal alignment, beyond what can be achieved through additional unstructured capacity in the projector.

A.6.2 The encoder reshapes the feature space

To probe how the encoder facilitates ICL, we quantified its effect on the geometry of modality-two class features (measured after the projector). With the encoder, the effective rank drops from 412 to 25, concentrating variance into a compact, structured subspace. This is accompanied by better cross-modal alignment with modality-one features: centered kernel alignment increases from 0.12 to 0.14 and the $L2$ distance decreases from 1.47 to 1.37. Together, these changes indicate that the encoder yields representations that are more compatible with the pretrained decoder's latent space.

A.6.3 The second modality encoder bottlenecks ICL

While synthetic data offers fine-grained control over distributional properties, it is significantly easier to classify compared to real-world inputs. To further investigate the role of the encoder under more realistic conditions, we extend our experiments to a multimodal setup where modality two consists of image data from the Omniglot dataset [7]. We train a series of encoders on Omniglot images with varying levels of difficulty, modulated via class noise, sample count, label distribution, and regularizations. These encoders achieve different levels of validation accuracy. For each pretrained encoder, we attached a projector and the same pretrained decoder from our unimodal experiments. We then fine-tuned the full model under various training schemes. As shown in Figure 8, we observed a clear trend: the higher the validation accuracy of the pretrained encoder, the stronger the resulting multimodal ICL capability. This finding aligns with empirical results from Zong et al. [18], who identified perception as a bottleneck for multimodal ICL. Moreover, we find that training only the encoder or encoder + projector results in ICL performance that quickly plateaus. In contrast, allowing the decoder to update during finetuning enables further improvements in ICL, suggesting that joint

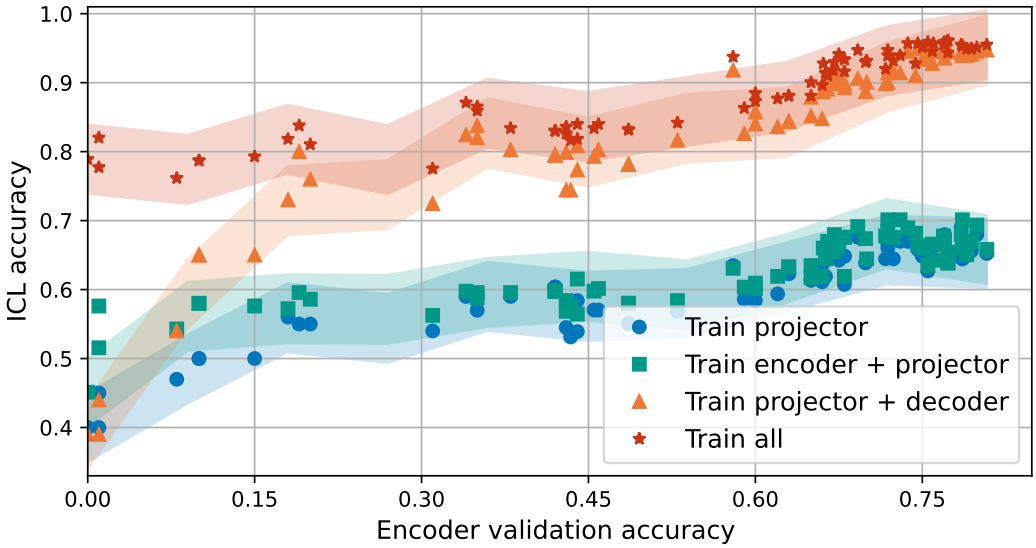

Figure 8: Encoder quality correlates with multimodal ICL on Omniglot. In this setup (Omniglot images as the second modality; decoder initialized from unimodal pretraining), ICL accuracy tends to increase as the pretrained encoder's validation accuracy improves. Saturation appears only when the decoder is not trained; allowing decoder updates continues to yield gains.

optimization across perception and reasoning modules is crucial for closing the gap in multimodal generalization.