PAIRED MULTIMODAL SCALING LAWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing multimodal scaling laws fit multimodality terms empirically after testing and never vary how much data is multimodally paired at fixed data budgets. We investigate how, under the same total data per modality, changing the number of paired data affects loss curves in multimodal classification tasks. We train models in three different environments and run experiment sweeps varying data sizes and pairing budget. Pairing ratios have a dramatic impact on loss and this impact is directly tied to how much information synergy the task contains. Only paired data is able to reduce synergistic loss, while unpaired data can reduce redundant or unimodal information up until unimodal floors. Unlike traditional scaling laws where loss drops immediately in power law decay, synergy acquisition is gated, requiring a critical threshold of paired data before synergistic loss falls at all. We introduce a new family of multimodal scaling laws where total data-attributable loss is the sum of four individual power laws corresponding to the four different information channels of redundancy, a unique channel per modality, and synergy, and show how this law is both more theoretically sound and empirically valid across our experiments. This law predicts multimodal loss in our experiments more accurately than existing laws, with 3.2% error on fit tests versus 10.4% error for the best pairing extension of published laws.

1 INTRODUCTION

Scaling laws predict how neural networks perform as a function of resources spent on it. This is traditionally represented in terms of model parameters N and training examples D (Hestness et al., 2017; Kaplan et al., 2020). Hoffmann et al. (2022) formulate this equation as

$$L(N, D) = \underbrace{E}_{\text{irreducible error}} + \underbrace{\frac{A}{N^{\alpha}}}_{\text{approximation error}} + \underbrace{\frac{B}{D^{\beta}}}_{\text{estimation error}}, \tag{1}$$

where the irreducible error represents the Bayes-optimal theoretical error floor, the approximation error is loss due to finite model size, and the estimation error is loss due to finite data sample.

Multimodal extensions of scaling laws keep the same shape as Hoffmann et al. (2022) but divide D into multiple subcomponents, i.e. by modeling loss as $L(N, D_1, D_2)$. Aghajanyan et al. (2023) decompose each modality into its own unimodal scaling law, then adjust empirically with multimodal interaction constants. Other laws represent loss in terms of D but with additional terms that capture the per-domain fraction of the total data (Ye et al., 2025; Tao et al., 2025). These laws represent multimodality by modeling *how much* of each modality is present.

However, a multimodal corpus contains a second quantity that is not modeled in existing scaling laws: **multimodal pairing**, or the presence of **both** modalities for a single data point. For an amount of paired data n_p, we represent the pairing fraction p as $p = \frac{2n_p}{D_1+D_2} \in [0, 1]$.

$p = 1$ only when every example is a pair and adding unpaired examples of either modality lowers p. Existing works have never varied p while holding D_1 and D_2 fixed, and we hypothesize that their failure to model p explicitly could partially explain why each work arrives at a different scaling law equation fitting their empirical results. When adding extra unpaired data has been tested, it is always in the text modality, and it always has come with the confound of adding extra total data, which means the effect of pairing alone has never been isolated.

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.

We run experiments that vary p across three settings at fixed data budget and find that p has a direct impact on downstream loss that existing laws fail to capture. We formulate new multimodal scaling laws that incorporate this variable based on Partial Information Decomposition (Bertschinger et al., 2014; Liang et al., 2023) and find our laws to be more theoretically grounded and empirically accurate. In particular, paired data is the only data stream that enables learning of **synergistic** information, or emergent information that arises only when both modalities are present. We also find unpaired data helps reduce non-synergistic information loss, but only with sufficient paired data alongside. Synergy is also gated: unlike traditional scaling law decay, synergistic loss stays at chance until a certain pair count threshold. Using these findings, we introduce a new four-channel multimodal scaling law, where data loss is the sum of four independent power law curves:

$$L(N, D_1, D_2, n_p) = E + \frac{A}{N^\alpha} + L_R + L_{U_1} + L_{U_2} + L_S \tag{2}$$

where each L term is fed data only by relevant modalities: redundant R by either modality, unique U_1 and U_2 by modalities 1 and 2, respectively, and synergistic S only by paired data n_p. This family of scaling laws fits our empirical results significantly better than existing multimodal scaling laws and are the only laws capable of modeling the synergistic phenomena we observe. Their form is also more grounded in multimodal interaction theory.

2 Related Work

2.1 Multimodal scaling laws

In unimodal settings (Eq. 1), while loss is a variable of N and D, A and B set the intercept of loss and the exponents α and β determine the rate of loss scaling. Exponents are set mainly by the task and the data distribution while architecture mainly moves the starting points A and B (Hestness et al., 2017; Kaplan et al., 2020). Aghajanyan et al. (2023) fits separate scaling curves per modality and bimodal runs to identify a synergy constant $C_{i,j}$ at odds with competition terms in N and D. Because synergy is modeled as a constant, a "competition barrier" exists at an N-D curve where synergy outweighs competition. Shukor et al. (2025b) train 457 image-text models across a variety of architectures and find little variation at scale but that early-fusion models perform better at lower parameter counts. Like Aghajanyan et al. (2023), this alludes to a "synergy threshold" at scale. Bill (Qingyun) Sun et al. (2024) propose weighting multimodal mixtures by tokenization efficiency and Tao et al. (2025) fit a compute-dependent modality mixture with a learned interaction matrix. Several of these works mix a text-only pool into paired image-text data, which lowers p but also adds data, so the role of p is never isolated.

2.2 Data-mixture scaling laws

Data mixture laws ask what mixture a corpus, typically text (web text, books, papers), should be assigned, either fit before training or updated online. Ye et al. (2025) present the law as minimization of individual domain losses and Liu et al. (2025) regress loss on a mixture of proxy models targeting 2% compute of a target run. Que et al. (2024) investigate the ratio of domains r as a scaling term and Shukor et al. (2025a) find mixture ratios that are fixed regardless of scale perform best. Other approaches operate online; Albalak et al. (2023) presents training as a bandit problem, and Jiang et al. (2025) calculate per-domain power laws during the run. Li et al. (2025) estimate loss bounds per-task via gradient norms. However, Chen et al. (2025) identify no method above consistently beats *uniform sampling* in general. They diagnose that all of the functional forms fail to model interaction terms between mixtures, so no equations generalize.

2.3 Partial information decomposition

Partial Information Decomposition, introduced by Williams & Beer (2010), decomposes the joint mutual information shared by two feature variables and a target variable into $I(X_1, X_2; Y) = R + U_1 + U_2 + S$, where R is redundant information shared between both modalities on Y, U_1 and U_2 are unique information per modality, and S is joint information present only with both modalities. Liang et al. (2023) formulate this problem for multimodal training loss. Using linear combinations of saturated unimodal training losses L_1 and L_2, multimodal loss $L_{1,2}$, and initial loss L, we are

	Digits		VQA-v2		NLVR2	
	image	audio	image	question	photos	sentence
Experimental details						
architecture	CNN + 4-way head (476k)		Frozen CLIP + head (5.3M)		Finetuned LXMERT (210M)	
chance loss L_0	$\ln 10 \approx 2.303$		$\ln 3129 \approx 8.048$		$\ln 2 \approx 0.693$	
D	8k, 16k, 32k, 64k, 128k, 256k		0, 4k… 256k		4k, 8k, 16k, 32k, 64k	
C ($k(D_1 + D_2)$ steps)	$k \in \{10, 20, 40\}$		$k \in \mathbb{Z}_+ \leq 80$		$k \in \{1, 2, 3\}$	
runs	408 (3 seeds)		3,660 (5 seeds)		1,010 (5 seeds)	
Unimodal data curve fitting $L(D) = E_N + B \cdot D^{-\beta}$ *at the above* N						
E_N	0.573	0.587	4.169	2.683	$\approx \ln 2$	$\approx \ln 2$
B	69.7	119	2.17	59.8	> 0	> 0
β	0.684	0.732	0.095	0.411	$\approx \infty$	$\approx \infty$
Partial information decomposition (nats)						
R	1.758		0.531		0	
U_1	1.758		0.127		0	
U_2	1.758		1.614		0	
S	1.758		0.144		0.532	

Table 1: Experimental details, unimodal data curve variables, and PID terms for each environment. Compute C represents number of training steps multiplied by batch size.

able to calculate $R + U_1$, $R + U_2$, and $S - R$, but not R or S in isolation. Bertschinger et al. (2014) and Liang et al. (2023) identify tractable ways to estimate S for joint distributions and multimodal machine learning. We assert that modeling S is the overlooked solution for better multimodal scaling laws. S determines how much co-occurring information is required for a task, which we hypothesize can only be learned during *paired* interaction. Meanwhile, we hypothesize *unpaired* data can help reduce U_1, U_2, and R independently but have **no impact** on S.

3 Experiments

We run experiments across three environments (Figure 1 and Table 1). We describe each experimental setup below with implementation details in Appendix B.

Digits. We construct series of four concatenated digits over MNIST images (LeCun et al., 1998) and AudioMNIST spectrograms (Becker et al., 2024) The label is a four-digit number corresponding directly to U_1, U_2, R, and S for each position: the first and second digits (U_1 and U_2) are the first and second digit shown in image and audio, the third digit (R) is the image and audio's third digit value, and the fourth digit (S) are the fourth input digits' sum $\mod 10$, so knowing only one modality is uninformative. We add symmetric label noise $\epsilon = 0.1$ so $U_1 = U_2 = R = S = \ln 10 - H_{\epsilon=0.1} \approx 1.758$. Loss is averaged over each label digit. The model architecture is two CNN encoders that pool to one value per position then a fused position-aware head.

VQA. We utilize VQA-v2 train, a collection of images and questions with a 3,129 label vocabulary. We freeze CLIP ViT-B/32 (Radford et al., 2021) image and caption embeddings. We then train a fusion head with unpaired images and captions accompanied with a learned null vector. As initially reported by Goyal et al. (2017), the VQA dataset is heavily dominated by the text alone.

NLVR2. NLVR2 is another image-question dataset engineered so multimodality is required. Each example is two photographs plus a sentence that asks a comparative question. All questions are binary with even splits between True/False, so individual unimodal accuracy hovers around 50%. We finetune an existing model, LXMERT Tan & Bansal (2019), which has achieved 74.45% on NLVR2, using the published model code[1].

3.1 Methodology

We apply the same methodology to each environment. We reserve a fixed, unseen test set for every run. Unpaired examples fill in the missing modality with a learnable null vector and a modality-

[1] https://github.com/airsplay/lxmert

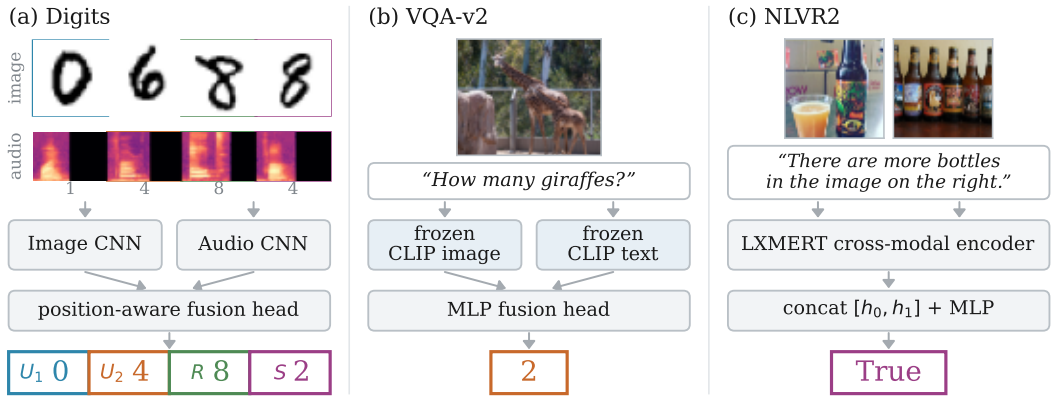


Figure 1: Illustration of each environment along with an example and tested architecture.

presence bit. During test time, both modalities are always present and loss is cross-entropy. In Appendix C.1, we analyze replacing cross-entropy with accuracy.

Grid sweeps and evaluation. We run sweeps across data budgets D, compute C, pairing count n_p, and multiple seeds. We fix architecture and parameter count N, as prior scaling law literature has demonstrated that loss due to architecture and parameter counts is independent of loss attributable to data composition (Hestness et al., 2017; Kaplan et al., 2020; Shukor et al., 2025b). For conciseness, we combine the irreducible and approximation errors E and A/N^{α} from Eq. 1 into a single term E_N. In Appendix C.2, we repeat experiments for Digits and VQA on different architectures and parameter sizes and again find the same qualitative results.

Counting paired and unpaired data. We determine total data via $D = D_1 + D_2$, i.e. the sum of unique examples per modality. This means as $p \to 1$, unique rows decrease. We do **not** split a pair into two unpaired data: each row is a unique label, so that improvement from $p = 0$ to $p = 1$ is tied to pairing and not label exposure. We fix equal compute via training step count, so a model trained at $D_{p=1}$ has twice as many training epochs as $D_{p=0}$ but the same gradient steps. In Appendix C.3, we consider other ways of counting paired and unpaired data.

Unimodal loss curves and PID. We run large-scale unimodal experiment runs until loss saturation and determine unimodal loss curves for each modality with Hoffmann-style variable estimates in Table 1. We use these curves to estimate PID variables R, U_1, U_2, and S. For Digits, this is analytically known to be ≈ 1.758 nats. For VQA, U_1, U_2, and R are estimated as linear combinations between the soft-label entropy $H(y) = 4.828$ nats minus the cross-entropy of saturated unimodal runs and a multimodal run, following Bertschinger et al. (2014). Shuffling questions within majority-answer class raises paired loss from 2.411 to 2.556 which estimates S as 0.144. NLVR2's unimodal probes always stay at chance, which we use to infer that 100% of reducible loss comes from synergy. We set $S = 0.532$ as the lower bound of mutual information based on the published human accuracy of 96.2% on the development split (Suhr et al., 2019).

4 How Does Pairing Affect Loss?

Paired data, and only paired data, is capable of reducing synergistic loss. Meanwhile, information conveyable in single modalities is learnable from unpaired data, but models require paired data to transfer unpaired learnings to a paired setting. In Figure 2, we keep total data per-modality fixed and equal ($D_1 = D_2$) and vary percentage of paired data. Increased pairing decreases loss in all three environments, and we hypothesize the relative percentage of reducible loss depends on task synergy after an initial one-time loss drop following $p = 0$. We identify this initial drop as pairs unlocking unpaired data: unpaired data is only useful in multimodal settings after learning paired interaction.

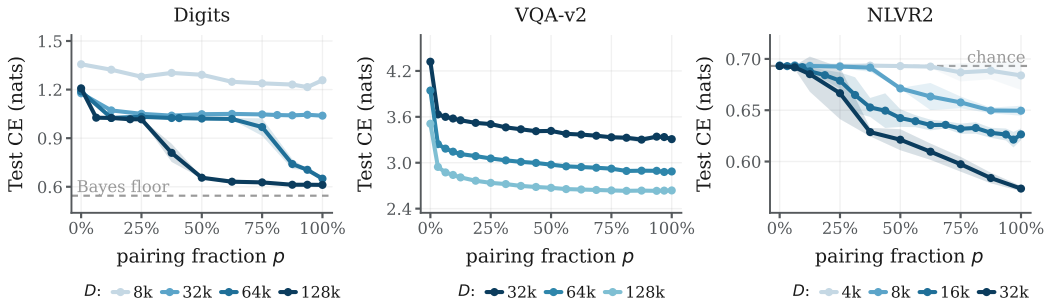


Figure 2: Test cross-entropy loss versus pairing fraction p at fixed per-modality data budgets $D_1 = D_2$ and fixed compute ($k \cdot D$ gradient steps; $k = 20$ for Digits and VQA, $k = 3$ for NLVR2). ± 1 std shown as shaded bands. For NLVR2, we filter out "failed-to-start" seeds, as discussed in §4.3.

In the Digits environment, for high D, we notice a **plateau then second descent**. This behavior reflects the model learning R, U_1, and U_2 labels very quickly and taking a longer time to learn S; a similar flatline then drop arises in NLVR2. Additionally, for NLVR2, some seeds never leave chance, with the likelihood of a "failed-to-start" seed increasing at lower p and lower D. We discuss these phenomena further in §4.3. In VQA, where synergy accounts for 6% of loss, there is a large loss dropoff from $p = 0\%$ to $p = 5\%$ but after a much more modest reduction in loss as $p \to 1$. At compute higher than $k = 20$ for VQA, optimal p lies interior of $p = 1$, which we note appears to be an artifact of overfitting because total examples halve as $p \to 1$. We conclude at fixed D, optimal compute decreases with increased p because repeat gradient steps per datum increase.

4.1 Controlling for Modality Presence Matching

One confound from our setup is that test examples always contain both modalities so improvement with p could be train-test alignment. This explains the jump in performance following $p = 0$ for Digits and VQA, because at $p = 0$ there are zero pairs to learn from. We re-run Digits runs with unpaired test examples. In Figure 3, we plot loss of each channel twice, on top with both modalities and on the bottom with irrelevant modalities nulled: for U_1, we null audio, for U_2, the image, and for R, both individually with the two losses averaged. When controlling for presence matching in this fashion, increasing pairing fraction has **no impact** on U_1 and U_2, and **slightly increases the loss** of R. We hypothesize that as $p \to 1$, the redundant prediction learns to rely on both channels simultaneously, so zeroing out a single channel actually hurts performance rather than having no effect like for U_1 and U_2. At low p, showing both modalities **increases loss** for single-modality channels (U_1, U_2, R) compared to just showing one modality because the model has not been exposed to enough paired examples in training to learn the other modality is irrelevant. After around 8k pairs, though, the top and bottom readouts converge to the same loss. We conclude, then, when controlling for train-test modality presence matching, paired data only helps reduce synergistic loss, but initially, pairing is required for unpaired training data to be useful during paired deployment for the model to learn that unimodal behavior can ignore irrelevant modalities.

4.2 When Does Extra Unpaired Data Help?

In Figure 4, we fix paired data and add extra unpaired examples. For Digits, we subdivide loss into the Bayes floor and each individual information channel's loss. Unpaired data reduces redundant and modality-unique loss but never synergistic loss. Additional unpaired data has the highest impact at lower overall D, with 24k more images or audio clips reducing loss at $D = 8$k from 1.26 to 1.10 nats; however, once the unimodal floors are more saturated at $D = 32$k, adding the same ratio barely affects loss. In both settings, synergistic loss is not reduced. This is consistent with NLVR2 seeing no improvement with additional unpaired data, since NVLR2 is dominated by S. Meanwhile, for VQA, adding unpaired text examples reduces loss more than images, which is consistent with relative PID weights for text and images. How much unpaired data helps also depends on the number of pairs; when $n_p = 128$, 124k unpaired questions reduce loss by 0.92 nats, but at $n_p = 4$k, it reduces loss by 1.20 nats. Pairs and unpaired data are therefore complements initially: adding some of one pool

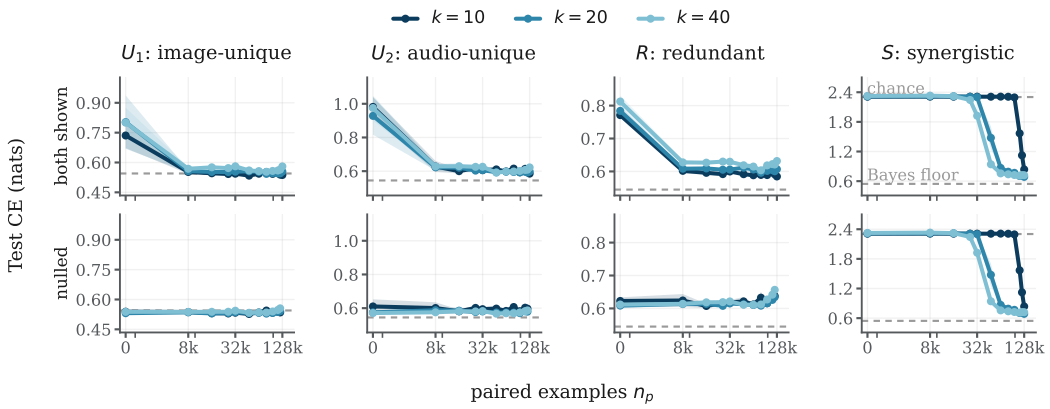


Figure 3: Individual channel losses at $D_1 = D_2 = 128k$ depending on which modalities are shown at test time. **Top:** loss when every test example shows both modalities. **Bottom** loss when irrelevant modalities are nulled for that given channel.

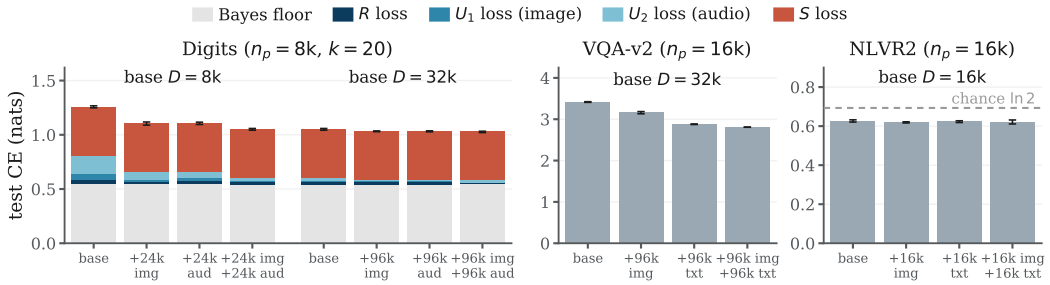


Figure 4: Test losses across the three environments at fixed n_p and k epochs, with additional unpaired data added. Digits loss is decomposed into individual information channel losses and the Bayes floor.

raises the value of the other. We use these results to conclude that **unpaired data can reduce loss but only to reduce R, U_1, and U_2 losses.**

4.3 Synergy Acquisition is Gated

Both the Digits S channel and the NLVR2 loss curves display atypical behavior at onset: they sit at chance initially then drop after a threshold of n_p. Meanwhile, VQA, whose synergistic loss share is small, and the R, U_1, and U_2 Digit channels, exhibit no such behavior and loss drops immediately. Synergy acquisition is gated on two factors: a minimum of paired examples n_{floor} and a minimum of paired exposures (including repeats) r. We propose a synergy gate $n^* = \max(n_{\text{floor}}, r/k)$.

For Digits, $n_{\text{floor}} \approx 2900$ and $r \approx 40k$ for an interleaved data mixture of paired and unpaired data. We find data presentation order matters. When paired data is entirely frontloaded, synergy is acquired after the paired data but forgotten by the end of training. When paired data is entirely backloaded, synergy is acquired, but at a worse rate: synergistic loss drops 0.60 nats during interleaved but only 0.15 nats during unpaired-then-paired training. In Figure 5, we display several properties of this gate. n^* is influenced by how much of the total loss is composed of synergy. As total loss share of S moves from 25% to 100%, n^* decreases: when S loss constitutes half of total loss, n^* is halved to 20k, and when S loss makes up 75%, n^* decreases to 12k. Synergy gating occurs for NLVR2, but unlike Digits, it is not a hard threshold gate but a random variable of n^*. NLVR2 random chance escape is best modeled as $\log_2 n^* \sim \mathcal{N}(11.32,\ 1.36^2)$, which models the actual escape distribution with median $n^* = 2552$ and a χ^2 fit of $p = 0.81$, with all seeds escaping at high n_p.

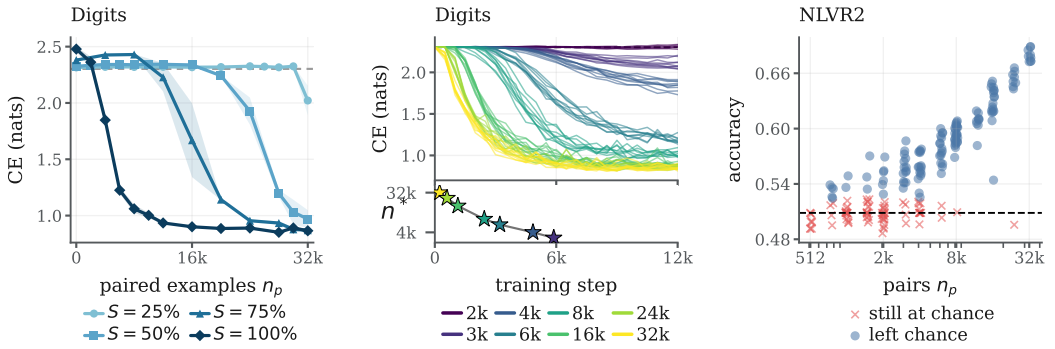


Figure 5: Evidence of a synergy gate. **(a)** Digits loss of the S channel as a factor of n_p and how much relative weight the S channel contributes to total loss at $D = 64k$ and $k = 40$. **(b) Top:** Digits loss of the S channel as a factor of n_p and training step (batch size$= 128$). **Bottom:** median training step and n^* per data budget when S loss leaves chance. n^* increases with higher data budgets because total repetitions lower. **(c)** NLVR2's "failed-to-start" seeds (red X's) versus seeds that leave chance accuracy (blue dots), plotted against n_p.

Law	k	Γ	$\partial\Gamma/\partial D \neq 0$	$L(N, D_1, D_2, n_p)$	at $p = 1$
Hoffmann Unimodal	4	≤ 0	✓	$E_N + B\, g(D_1{+}D_2)$	$E_N + B\, g(2D)$
Hoffmann on unique D	4	≤ 0	✓	$E_N + B\, g(D_{\text{tot}})$	$E_N + B\, g(D)$
Hoffmann on D_1 and D_2	7	$= 0$	×	$E_N + B_1 g_1(D_1) + B_2 g_2(D_2)$	$E_N + B_1 g_1(D) + B_2 g_2(D)$
Aghajanyan $+Cp$	8	$= 0$	×	$E_N + B_1 g_1(D_1) + B_2 g_2(D_2) + Cp$	$E_N + B_1 g_1(D) + B_2 g_2(D) + C$
Shukor mixture	6	< 0	✓	$E_N + (C_1 p + C_2(1{-}p))^{-1} + B\, g(D_{\text{tot}})$	$E_N + C_1^{-1} + B\, g(D)$
Ye mixture	7	any	✓	$E_N + C \exp(\alpha_2 n_2/D_{\text{tot}} + \alpha_p n_p/D_{\text{tot}}) + B\, g(D_{\text{tot}})$	$E_N + C\, e^{\alpha_p} + B\, g(D)$
Four-channel family	k / Δk			*base law and switch substitutions*	*what it models*
Base	14	≤ 0		$E_N + \sum_a B_a\, g_a(D_a)$	one curve per PID channel
Shared	-6	no change		$g_a \to g$	one (δ, β) for all channels
Gate	$+0$	any	✓	$B_S\, g_S(n_p) \to B_S/(1{+}(n_p/n_\star)^\gamma)$	synergy is gated
PID	-3	no change		$B_a \to s_0 W_a$	amplitudes from measured PID
Effective Count	$+2$	any		$D_i \to D_i - (1{-}\lambda_0)\, n_i\, \frac{n_\lambda}{n_p + n_\lambda}$	pairs unlock unpaired data

Table 2: Extensions of existing scaling laws (Top) and the introduced four-channel law and its variants (Bottom). Existing laws reduce to original equations at $p = 1$. Each of the variations of the four-channel law can be applied independently for 16 combinations. k is the number of fitted parameters for the corresponding law, $D_{\text{tot}} = D_1 + D_2 - n_p$, $n_i = D_i - n_p$, and W_a are measured PID weights. $\Gamma = -\partial^2 L/\partial D_1 \partial D_2$ is measured at p. Γ and $\partial\Gamma/\partial D$ are discussed in §5.2. We use $g(x) = (\delta/(D + \delta))^\beta$.

5 Paired Scaling Laws

We now attempt to form a generalizable multimodal scaling law with pairing as an input parameter, $L(N, D_1, D_2, n_p)$. We propose several extensions of existing scaling laws by adding p as a variable where at $p = 1$ they simplify to the original equation. We then introduce a new loss family, where data loss is summed as four separate loss curves based on R, U_1, U_2, and S, and permute four law variations of this family. These new laws consistently outperform pairing-variable extensions of published laws, even at $p = 1$. We provide theoretical backing for this (namely, that existing laws cannot model synergy) and show how these laws can find optimal pairing ratios under a pairing budget where pairs are more expensive than unpaired data.

Extending existing laws. We provide six candidates from existing laws (Table 2, Top) and extend their forms to model pairing. The first three are traditional Hoffmann et al. (2022)-style equations with different approaches of counting data: as the total data exposures per modality, the number of unique data rows, and two independent data curves per modality. We then extend an existing multimodal scaling law and two data mixture laws from the literature: Aghajanyan et al. (2023)'s law that adds a constant "synergy" factor, Shukor et al. (2025a)'s law that provides a "usefulness" term to each mixture (paired and unpaired), and Ye et al. (2025)'s exponential of the mixture proportions.

	Law	Shared	Gate	PID	Eff.	k	Digits $D\uparrow$	Digits $p{=}1$	Digits R^2	VQA $D\uparrow$	VQA $p{=}1$	VQA R^2	NLVR2 $D\uparrow$	NLVR2 $p{=}1$	NLVR2 R^2	Overall $D\uparrow$	Overall $p{=}1$	Overall R^2
	Hoffmann Unimodal					4	13.9%	30.4%	0.00	8.2%	8.1%	0.00	1.8%	6.6%	0.00	8.0%	15.0%	0.00
	Hoffmann on unique D					4	14.9%	37.1%	-0.12	10.5%	13.1%	-1.26	1.7%	6.9%	-0.45	9.0%	19.0%	-0.61
	Hoffmann on D_1 and D_2					7	11.5%	27.3%	0.00	2.7%	6.7%	0.00	1.8%	6.6%	0.00	5.3%	13.5%	0.00
	Aghajanyan $+Cp$					8	11.5%	20.4%	0.39	2.3%	5.0%	0.40	1.5%	6.0%	0.22	5.1%	10.4%	0.34
	Shukor mixture					6	9.6%	29.1%	0.46	8.2%	9.4%	0.02	1.4%	6.2%	-0.09	6.4%	14.9%	0.13
	Ye mixture					7	9.9%	31.8%	0.46	5.3%	8.1%	0.06	1.3%	5.7%	0.11	5.5%	15.2%	0.21
Four-channel		×	×	×	×	14	9.6%	21.4%	0.63	3.0%	3.4%	0.80	**0.8%**	4.7%	0.61	4.5%	9.9%	0.68
		×	×	×	✓	16	9.5%	9.9%	0.75	3.2%	**1.8%**	**0.86**	**0.8%**	4.8%	0.61	4.5%	5.5%	0.74
		×	✓	×	×	14	9.5%	21.7%	-1.27	2.9%	3.0%	0.81	1.1%	2.8%	**0.81**	4.5%	9.2%	0.12
		×	✓	×	✓	16	9.5%	**3.3%**	**0.95**	3.8%	3.5%	**0.86**	1.1%	3.0%	**0.81**	4.8%	3.3%	**0.87**
		✓	×	×	×	8	9.2%	23.9%	0.66	2.9%	2.1%	0.80	**0.8%**	4.7%	0.61	4.3%	10.2%	0.69
		✓	×	×	✓	10	9.2%	23.9%	0.74	2.9%	2.4%	0.84	**0.8%**	4.7%	0.61	4.3%	10.3%	0.73
		✓	✓	×	×	10	9.2%	21.6%	-1.56	2.9%	2.8%	0.81	1.1%	2.7%	**0.81**	4.4%	9.0%	0.02
		✓	✓	×	✓	12	9.3%	**3.3%**	**0.95**	3.2%	3.5%	0.85	1.1%	2.8%	**0.81**	4.5%	**3.2%**	**0.87**
		×	×	✓	×	11	**8.6%**	21.5%	0.63	2.3%	2.7%	0.81	**0.8%**	4.6%	0.61	**3.9%**	9.6%	0.68
		×	×	✓	✓	13	9.0%	21.5%	0.74	**2.1%**	2.2%	**0.86**	**0.8%**	4.6%	0.61	4.0%	9.4%	0.74
		×	✓	✓	×	11	9.8%	21.4%	-1.10	2.4%	3.1%	0.79	1.1%	**2.1%**	**0.81**	4.4%	8.9%	0.17
		×	✓	✓	✓	13	9.8%	8.5%	**0.95**	2.6%	3.5%	**0.86**	1.1%	**2.1%**	**0.81**	4.5%	4.7%	**0.87**
		✓	×	✓	×	5	10.4%	22.5%	0.49	3.2%	6.2%	0.22	**0.8%**	4.6%	0.61	4.8%	11.1%	0.44
		✓	×	✓	✓	7	10.6%	22.5%	0.54	2.8%	4.7%	0.61	**0.8%**	4.6%	0.61	4.7%	10.6%	0.59
		✓	✓	✓	×	7	9.8%	23.3%	0.63	3.0%	3.2%	0.79	1.1%	**2.1%**	**0.81**	4.6%	9.5%	0.74
		✓	✓	✓	✓	9	9.8%	5.7%	**0.95**	2.8%	3.7%	0.83	1.1%	**2.1%**	**0.81**	4.6%	3.8%	0.86

Table 3: Paired scaling law scores on the three fit tests: percent error for $D \uparrow$ and $p = 1$ (lower is better), and within-cell R^2 (higher is better).

The four-channel law. Building on our analysis that R, U_1, U_2, and S explains much of the variation between environments, we present an alternative equation family, that **total loss can be broken down as the sum of independent power law curves for each information channel**:

$$L(N, D_1, D_2, n_p) = E_N + L_R + L_{U_1} + L_{U_2} + L_S \tag{3}$$

Each individual L term is a data power curve where the total amount of "data" for that term corresponds to the number of examples relevant to its channel. R is fed by $D_1 + \kappa D_2$, where κ is the learned relative usefulness per modality, and U_1, U_2 and S are fed by D_1, D_2, and n_p. As we fix architecture and model size in our experiments, $E_N = E + A/N^\alpha$ is a constant. We vary four factors across this equation family for a total of sixteen variations:

1. Shared δ and β: whether the data curve intercepts and exponents are shared across the four losses or independent per curve
2. Synergy offset: whether the synergy gating effect is modeled in L_S via a soft gate $1/(1 + (n_p/n^*)^\gamma)$, based on §4.3
3. PID-pinned: whether loss terms' relative weights are set by the environment's PID ratios
4. Effective counts: whether unpaired example count is discounted until a critical threshold of paired data, based on the "pairs unlock unpaired data" finding from §4.1

We replace B/D^β with $(L_0 - E_N)\big(\delta/(D + \delta)\big)^\beta$, which converge at $D >> \delta$, for better-behaved curves as $D \to 0$ and let $g(D) = (\delta/(D + \delta))^\beta$ for conciseness. Full derivations in Appendix A.

5.1 Evaluation

We evaluate each law based on four held-out tests after fitting each equation via least squares to each test's respective training runs. D_{max} represents the size of data for the largest run of each environment. The three tests are:

1. **Scale** ($D\uparrow$): fit on runs with $D \in [0, D_{\text{max}}/2)$ and test on runs with $D \in [D_{\text{max}}/2, D_{\text{max}}]$
2. **Full pairing** ($p = 1$): fit on runs where $p \in [0, 1/2)$ and test on the corner $p = 1$
3. **Pairing shape** (R^2): fit on all except one (D_1, D_2) grid, predict every loss for every p at (D_1, D_2), subtract the mean, then measure explained variance

5.2 Results

Results of each fit are in Table 3. Extensions of published laws are the weakest fits across all tests and environments, with every variant of the four-channel law beating published laws on $D \uparrow$, 14 of

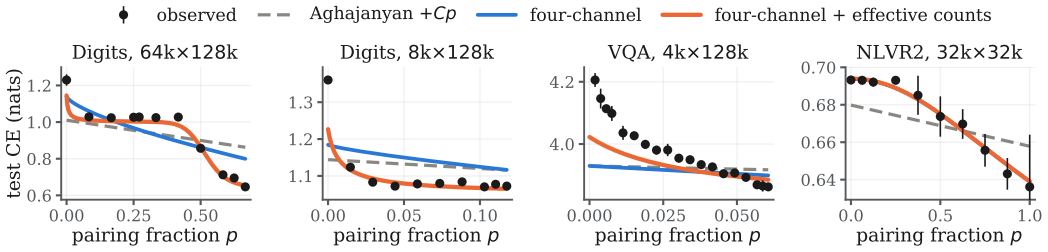


Figure 6: Paired scaling laws on four different D grids for the Leave-One-Out test, meaning each curve is fit on other (D_1, D_2) combinations and tested on this (D_1, D_2) cell. All cells, laws, and tests can be viewed at this web link[2] or in the Supplementary Materials.

16 on $p = 1$, and 13 of 16 on R^2. The two factors that are most helpful for the four-channel law are the synergy gate and effective counts. Shared exponents reduce the number of fitted parameters and also improve overall fit. PID loss weights do not improve test performance, although they do reduce the number of fitted parameters.

Theory behind the four-channel law's improvement. We identify two theoretical reasons why the four-channel law outperforms extensions. Let $\Gamma = -\partial^2 L / \partial D_1 \partial D_2$ at fixed p, which represents how much adding an example of modality 1 changes the value of adding an example of modality 2. Properly modeling synergy requires $\Gamma > 0$ at some point, since synergistic information can only be acquired when both modalities are present. Additionally, proper modeling requires $\partial\Gamma/\partial D$ must be non-zero, i.e. Γ changes with changing data, as different information channels eventually saturate. Every existing law except the Ye mixture violates either or both of these constraints (Table 2), and Ye only models relative share instead of absolute counts as observed in §4. The four-channel law contains both properties via the S decomposition, where synergy loss is gated before a threshold (allowing $\Gamma > 0$) and decays (allowing $\partial\Gamma/\partial D \neq 0$).

Interactive viewer. We display examples of different law fits plotting CE against p in Figure 6. Four-channel laws are able to successfully model synergy gates, delayed loss, and p-dependent loss. To further illustrate these results, we have included an interactive paired scaling law viewer in the Supplementary Materials and at this public web link[2].

Optimal pairing fractions. These laws can also be used to determine the best pairing fraction under a budget where pairs are expensive. For NLVR2, which is almost all synergy, optimal pairing is always $p = 1$, but for Digits and VQA, it moves well below $p = 1$ when pairs cost two to three times unpaired data (Appendix A.5).

6 Conclusion

We investigated how multimodal loss changes as a function of data pairing and found that only pairs can reduce synergistic loss while unpaired data can reduce redundant and modal-unique loss. We introduced a new four-channel multimodal scaling law that outperforms existing scaling laws and also proved that data pairing is an independent variable that has measurable impact on model loss.

Scope and limitations. Partial information decomposition is defined with respect to a label Y, so our laws are most applicable to supervised multimodal tasks. We did not investigate self-supervised generative pretraining without labels where pairing itself is the training signal. We leave investigation of the generalizability of our four-channel scaling law in self-supervised training regimes as future work, as well as larger data and parameter scales. We note many established scaling law papers have been tested at scales comparable to our experiments (Hestness et al., 2017; Rosenfeld et al., 2020; Sorscher et al., 2022; Bahri et al., 2024) and that the power-law forms they describe hold at larger scales (Kaplan et al., 2020; Hoffmann et al., 2022; Aghajanyan et al., 2023).

[2] https://paired-scaling-laws-viewer.up.railway.app

AI USE STATEMENT

In this work, we performed the following tasks **without** the help of AI or LLM usage:

- Formulating core mathematical claims and theorems
- Developing theoretical models and conceptual frameworks
- Proposing and refining hypotheses
- Discovering relevant research areas and important existing literature
- Designing research methodology and experiments
- Interpreting results
- Identifying experimental parameters
- Verifying validity of claims
- Designing the paper structure
- Crafting the title and keywords
- Writing the final manuscript

The following tasks were performed **with** the help of AI and LLM usage, with brief justification:

- Providing critical ingredients and writing of proofs: We used AI and LLMs to double-check and verify the proposed proofs as well as simulate different variations of each equation and check for edge cases.
- Providing feedback for research methodology and experiments: After outlining the core experiments, we used AI to help brainstorm potential design flaws and ablations. A human was always present to review and approve ablations.
- Implementing methods and creating code: We used AI and LLMs to generate the code to execute our experiments. The majority of our code is AI-written, including the pairing viewer, but the authors have reviewed the code manually and ran unit tests, sanity checks, and manually validated that the execution and results are error-free as best to the authors' ability.
- Creating or modifying figures or images: We used AI and LLMs to generate figures and tables, mostly for ease of formatting. All ideas and layouts for figures and tables came from the authors directly. The results and layout were then manually adjusted and refined by the authors.
- Cleaning and reformating datasets: We used LLMs to standardize our several datasets in the codebase.
- Creating draft sections of the paper: We used AI and LLMs to summarize and create draft sections of several parts of the paper; however, every sentence in the final manuscript is human-written.
- Performing related literature sweeps: After the key relevant research areas were identified through manual review, we used AI and LLMs to exhaustively search for all relevant literature.
- Grammar and formatting: We used AI and LLMs to perform final grammar and formatting checks in the manuscript.

The following tasks are not applicable to this work:

- Generating synthetic data sets
- Assisting with translation
- Supporting qualitative and thematic data analysis
- Formulating questions for surveys or interviews
- Transcribing recordings

We have reviewed all AI-assisted work and take full responsibility for the final content of this work.

Ethics Statement

We do not believe this work raises any ethical concerns, as this work is solely focused on fundamental machine learning theory research and empirical analysis. This work does not collect new data, deploy a system, or present any potential data privacy concerns. We use existing research datasets in compliance with their usage licenses. In our Supplementary Material, we only provide new code and run logs and do not redistribute external dataset images and text.

Reproducibility Statement

We are fully committed to the scientific pillar of reproducibility and transparency. We provide additional written detail of experimental implementation in Appendix B. We also have included the source code for all of our experiments in the Supplementary Material.

Running all of the experiments for this work required significant computational resources; by our estimates, the totality of all of our runs took about 4000 GPU-hours and $3e19$ FLOPs. These runs were conducted on a high-performance computer cluster where dozens of P100 GPUs were run in parallel over the course of several months. While we have provided the code necessary to re-run these experiments, we also separately provide the logs and results of our runs in a separate folder in the Supplementary Material and as an anonymous online pairing law viewer at this link so that future work does not need to spend unnecessary resources to access our results.

Due to the high GPU-hour demand that a full parameter sweep requires, some of the runs that appear only in the pairing law viewer were only run for a single seed. However, all training runs that were discussed and included in the figures of the main paper were run for multiple seeds as per Table 1.

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 265–279. PMLR, 2023. URL https://proceedings.mlr.press/v202/aghajanyan23a.html.

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. In *R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023*, 2023. arXiv:2312.02406.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015. doi: 10.1109/ICCV.2015.279.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121.

Sören Becker, Johanna Vielhaben, Marcel Ackermann, Klaus-Robert Müller, Sebastian Lapuschkin, and Wojciech Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024. ISSN 0016-0032. doi: 10.1016/j.jfranklin.2023.11.038. URL https://www.sciencedirect.com/science/article/pii/S0016003223007536.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. doi: 10.3390/e16042161.

Bill (Qingyun) Sun, Gavin (Zhen) Guo, and PIN AI Team. Scaling law hypothesis for multimodal model, 2024. URL https://arxiv.org/abs/2409.06754.

Mayee F. Chen, Michael Y. Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=sZGZJhaNSe. arXiv:2411.05735.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. doi: 10.1109/CVPR.2017.670.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically, 2017. URL https://arxiv.org/abs/1712.00409.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=aqok1UX7Z1. arXiv:2410.11820.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361. Preprint; no formal publication found.

Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. The MNIST database of handwritten digits. Online, 1998. URL http://yann.lecun.com/exdb/mnist/. Dataset. The accompanying paper is LeCun, Bottou, Bengio, and Haffner, Proceedings of the IEEE, 86(11):2278–2324, 1998.

Zeman Li, Yuan Deng, Peilin Zhong, Meisam Razaviyayn, and Vahab Mirrokni. Pike: Adaptive data mixing for large-scale multi-task learning under low gradient conflicts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 187600–187640, 2025. doi: 10.52202/085713-5640. arXiv:2502.06244.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27351–27393, 2023. doi: 10.52202/075280-1192.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=5BjQOUXq7i. arXiv:2407.01492.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. D-cpt law: Domain-specific continual pre-training scaling law for large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 90318–90354, 2024. doi: 10.52202/079017-2867. arXiv:2406.01375.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryenvpEKDr.

Mustafa Shukor, Louis Béthune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. In *Advances in Neural Information Processing Systems*, volume 38, pp. 143691–143716, 2025a. doi: 10.52202/085713-4314. arXiv:2507.09404.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12–23, 2025b. doi: 10.1109/ICCV51701.2025.00009.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19523–19536, 2022. doi: 10.52202/068431-1419.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514/.

Chaofan Tao, Zhuo Li, Xiao-Hui Li, Jierun Chen, Weike Jin, Haoli Bai, Lifeng Shang, and Lu Hou. Modalmix: Optimizing multimodal data mixtures with compute-dependent regression. OpenReview, 2025. URL https://openreview.net/forum?id=R1HcIN90A1. Submitted to ICLR 2026; not in the ICLR 2026 proceedings.

Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information, 2010. URL https://arxiv.org/abs/1004.2515. Preprint; no formal publication found.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=jjCB27TMK3. arXiv:2403.16952.

A PAIRED SCALING-LAW DERIVATIONS

In this appendix section, we derive δ (A.1) and Γ (A.2), define effective counts (A.3), show two additional law fit results (A.4), and plot optimal pairing ratios under a budget (A.5).

A.1 SINGLE-SOURCE LOSS CURVES WITH δ

The Hoffmann form (Eq. 1) diverges as $D \to 0$ but classification loss is bounded by the loss of an untrained model, L_0 ($\ln k$ for k balanced classes). To make this equation well-behaved at low D, we add a normalization term δ, which acts as a pseudocount of seen examples:

$$L(D) = E_N + \frac{B}{(D+\delta)^\beta}. \tag{4}$$

At $D = 0$ this gives $L_0 = E_N + B/\delta^\beta$, which rewritten yields $B = (L_0 - E_N)\,\delta^\beta$ and

$$L(D) = E_N + (L_0 - E_N)\,g(D), \qquad g(D) = \Big(\frac{\delta}{D+\delta}\Big)^\beta. \tag{5}$$

For $D \gg \delta$ this approaches the original term B/D^β. Both endpoints L_0 and E_N can be directly measured via expected loss of a uniform distribution and a single long training run that saturates on data and compute, respectively. This leaves the exponents δ and β to be fit instead of B and β (as B can derived through L_0, E_N, δ, and β). We can thus estimate the value of a single example at a given loss:

$$\frac{\mathrm{d}L}{\mathrm{d}D} = -\frac{\beta}{D+\delta}\,\big(L(D) - E_N\big). \tag{6}$$

Jiang et al. (2025) use a similar quantity, without δ, to choose data mixtures during training.

A.2 INTERACTION SIGN Γ

Let $v_i = -\partial L/\partial D_i$ be the value of adding one more example of modality i. This value is always positive as loss curves are monotonically decreasing, and comparing v_i to v_j determines whether modality i or j is more locally helpful. We define an interaction sign Γ as

$$\Gamma = \frac{\partial v_i}{\partial D_j} = -\frac{\partial^2 L}{\partial D_i\,\partial D_j}. \tag{7}$$

If $\Gamma > 0$, the modalities are complements: more of j makes an example of i more valuable. If $\Gamma = 0$, they are independent, and if $\Gamma < 0$, they are substitutes. A multimodal law that can express synergy should be able to express all three, as there are scenarios when one modality helps, hurts, or is neutral to additional data from the other modality.

Proof that existing laws (except Ye) have $\Gamma \le 0$ at fixed p. All Hoffmann-style equations take the form $L = E_N + W g(D_i + \kappa D_j)$ with $\kappa \ge 0$. g is a decreasing convex curve, thus

$$\frac{\partial^2 L}{\partial D_i\,\partial D_j} = W\kappa\, g''(D_i + \kappa D_j) \ge 0 \qquad \Longleftrightarrow \qquad \Gamma \le 0 \tag{8}$$

for every parameter setting. This means Hoffmann-derived equations model modalities i and j as substitutes only. Similarly, at fixed p, Aghajanyan's Cp and Shukor's $(C_1 p + C_2(1-p))^{-1}$ are constants, so Aghajanyan's Γ is 0, and Shukor's is negative from its pooled $g(D_{\mathrm{tot}})$. Ye's exponent depends on the share $(D_2 - n_p)/D_{\mathrm{tot}}$, which at fixed p can change because n_p is free even at fixed p by allocating additional D_1.

Proof that the four-channel law does not constrain Γ at fixed p. The four-channel family is

$$L = E_N + L_R + L_{U_1} + L_{U_2} + L_S, \tag{9}$$

$$D_R = D_1 + \kappa D_2, \quad D_{U_1} = D_1, \quad D_{U_2} = D_2, \quad D_S = n_p. \tag{10}$$

	Law				k	Digits		VQA		NLVR2		Overall	
						LOO	Full	LOO	Full	LOO	Full	LOO	Full
	Hoffmann Unimodal				4	5.9%	5.6%	3.9%	3.7%	1.0%	1.3%	3.6%	3.5%
	Hoffmann on unique D				4	5.2%	4.5%	4.0%	3.8%	1.0%	1.2%	3.4%	3.1%
	Hoffmann on D_1 and D_2				7	4.2%	3.6%	1.1%	1.0%	1.2%	1.3%	2.2%	2.0%
	Aghajanyan $+Cp$				8	5.7%	4.6%	1.2%	1.0%	0.9%	0.8%	2.6%	2.2%
	Shukor mixture				6	4.1%	3.9%	4.6%	4.3%	0.8%	1.1%	3.2%	3.1%
	Ye mixture				7	4.4%	4.1%	2.0%	1.7%	0.9%	1.0%	2.4%	2.3%
Four-channel	Shared	Gate	PID	Eff.									
	×	×	×	×	14	4.5%	4.2%	0.8%	**0.6%**	0.5%	0.5%	1.9%	1.8%
	×	×	×	✓	16	2.8%	2.6%	**0.7%**	**0.6%**	0.5%	0.5%	1.4%	1.2%
	×	✓	×	×	14	3.2%	4.2%	0.8%	0.7%	**0.2%**	**0.3%**	1.4%	1.7%
	×	✓	×	✓	16	**1.7%**	**1.4%**	0.8%	**0.6%**	**0.2%**	**0.3%**	0.9%	**0.8%**
	✓	×	×	×	8	5.2%	4.9%	0.8%	0.7%	0.5%	0.5%	2.1%	2.0%
	✓	×	×	✓	10	4.1%	3.7%	**0.7%**	**0.6%**	0.5%	0.5%	1.8%	1.6%
	✓	✓	×	×	10	3.3%	3.9%	0.8%	0.7%	**0.2%**	**0.3%**	1.4%	1.6%
	✓	✓	×	✓	12	**1.7%**	**1.4%**	**0.7%**	**0.6%**	**0.2%**	**0.3%**	**0.8%**	**0.8%**
	×	×	✓	×	11	4.6%	4.3%	0.8%	**0.6%**	0.5%	0.5%	2.0%	1.8%
	×	×	✓	✓	13	3.5%	3.1%	**0.7%**	0.7%	0.5%	0.5%	1.6%	1.4%
	×	✓	✓	×	11	3.3%	4.3%	0.8%	0.7%	**0.2%**	**0.3%**	1.4%	1.8%
	×	✓	✓	✓	13	2.7%	1.6%	**0.7%**	**0.6%**	**0.2%**	**0.3%**	1.2%	**0.8%**
	✓	×	✓	×	5	3.4%	2.9%	1.4%	1.2%	0.5%	0.5%	1.7%	1.5%
	✓	×	✓	✓	7	4.8%	4.7%	0.9%	0.8%	0.5%	0.5%	2.1%	2.0%
	✓	✓	✓	×	7	3.5%	2.8%	**0.7%**	0.7%	**0.2%**	**0.3%**	1.5%	1.3%
	✓	✓	✓	✓	9	2.3%	1.9%	**0.7%**	0.7%	**0.2%**	**0.3%**	1.1%	0.9%

Table 4: Paired scaling law scores of the two interpolation tests: Leave-One-Out and the Full test.

With power-law channels, $L_a = B_a g(D_a)$, every term is a convex function of one count or of a sum. The base law therefore has $\Gamma \leq 0$ like the Hoffmann forms. However, adding the synergy gate replaces L_S with $B_S/(1 + (n_p/n_\star)^\gamma)$. For $\gamma > 1$ this curve is concave below $n_p = n_\star\big((\gamma - 1)/(\gamma + 1)\big)^{1/\gamma}$, so there L_S contributes $-(p/2)^2 L_S'' > 0$ to Γ. Before synergy is learned, more of either modality makes the other more valuable, because together they buy the pairs that open the gate. After the gate, the curve is convex again and the Γ becomes negative. Effective counts change the counts that feed each channel and can in rare cases make Γ positive.

A.3 Effective counts

A model trained without pairs can model unimodal channels effectively but see increased loss during test-time paired data (§4.1). We model this as unpaired data being less valuable until a certain amount of paired data is available to align paired data with unpaired representations. At zero pairs, an unpaired example is worth $\lambda_0 \leq 1$ of a paired one and effective count is represented as:

$$D_i^{\text{eff}} = D_i - (1 - \lambda_0)\, n_i \, \frac{n_\lambda}{n_p + n_\lambda}. \tag{11}$$

$n_i = D_i - n_p$ is the unpaired count of modality i and n_λ is the number of pairs that unlocks half the locked value. More pairs increases the effective counts of unpaired data, and thus pairs and unpaired examples become complements initially.

A.4 Additional Law Fit Tests

In Table 4, we report two interpolation law fits: the Leave-One-Out test, where an equation curve is fit on all cells except one (D_1, D_2) grid and then tested on every p of that heldout (D_1, D_2) cell, and the Full test, where no cells are heldout and one law is fit over every data point. These two tests reveal the same results as the main paper, that the four-channel law outperforms extensions of existing laws.

A.5 Choosing a Pairing Fraction Under a Budget

We consider the practical problem of the optimal pairing ratio under a resource budget where pairs cost more than unpaired data. Figure 7 plots loss versus p given a pair price $\phi \in \{1, 2, 3\}$ for the three environments. When $\phi = 1$ (i.e. pairs do not cost extra), optimal $p = 1$, but higher ϕ changes p depending on the environment: for NLVR2, which is dominated by synergy, optimal p is always 1, but for Digits and VQA, this value lies in the interior.

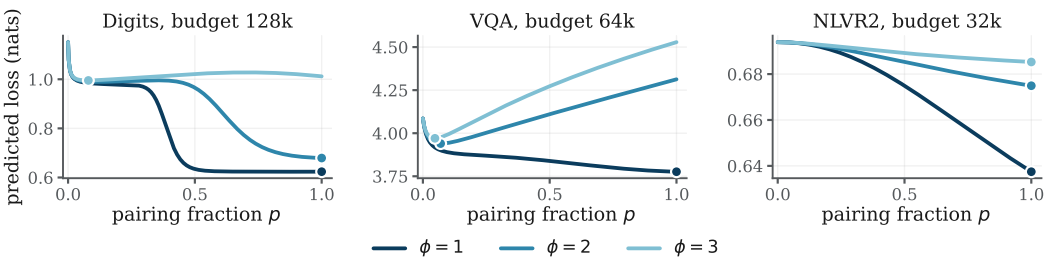


Figure 7: Predicted loss versus pairing fraction p at pair prices $\phi \in \{1, 2, 3\}$. Dots mark optima.

B Additional Experiment Implementation Details

In the sections below, we provide additional implementation detail for each of the environments.

B.1 Digits Experiment

Labels and known constants. Each modality shows four digits, one per label position (§3). The image is four MNIST digits side by side ($1 \times 28 \times 112$). The audio is four AudioMNIST log-mel spectrograms joined in time ($1 \times 64 \times 400$). Training audio comes from speakers 1–48 and test audio from speakers 49–60. Test images come from the MNIST test split. Symmetric label noise at rate $\varepsilon = 0.1$ is applied to the label only. Output heads are initialized to zero so that an untrained model predicts the uniform distribution.

Model. For each modality, we use the same encoder: three blocks of a 3×3 convolution (32, 64, and 128 channels and max-pooling), then GroupNorm and ReLU. The encoder pools to a single column per digit position and then projects to a 128-d embedding. We train with auxiliary cross-entropy of weight 1 for modality presence to aid with modular sum for the synergy digit. Missing modalities are replaced with a null embedding. The fusion trunk is a two-layer neural network with hidden layer width of 256, which sees both embeddings and the modality presence bits, then feeds into four ten-way heads, one for each label digit. This model has $483k$ trainable parameters.

Training and evaluation. We used AdamW with learning rate 3×10^{-4} and weight decay 10^{-2}, batch size 128 drawn uniformly with replacement, cosine decay over the planned number of steps, and gradient clipping at 2. We hold out one tenth of training for a validation set, used for logging. We trained for $k \in \{10, 20, 40\}$ epochs. The test set is 5000 examples.

B.2 VQA-v2 Experiment

Data. The train split of VQA-v2 has $443,757$ questions on COCO train204 images with 10 annotator answers per question (Antol et al., 2015; Goyal et al., 2017). The vocabulary is the 3,129 answers that occur at least 64 times. The 8,872 questions with no answer in the vocabulary are omitted, leaving 434,885 questions on 82,772 images. The target label is the soft label of annotator votes, and loss is the cross-entropy in nats against the soft label. The test set is 25,000 val2014 questions.

Unpaired examples. We define an unpaired image as an image with no text, with the answer label sampled independently from any of the questions that are asociated with that image. An unpaired text example is a normal question example with the image nulled.

Model. We used CLIP ViT-B/32 (Radford et al., 2021) to embed images and questions one time into a frozen cache. Using these embeddings, we trained an MLP head ($1026 \to w \to w \to 3129$ then GELU) over the two embeddings and two presence bits. In the main experiment, $w = 1024$ (5.3M parameters), and in Appendix C.2 we test $w = 512$ and 256 (2.39M and 1.13M). The last layer is zero-initialized. We used AdamW, learning rate 3×10^{-4}, weight decay 0.01, batch size 512. One sixteenth of training data was heldout as validation.

B.3 NLVR2 Experiment

Data. We used the LXMERT release of NLVR2 train, which contains $86{,}373$ examples (Suhr et al., 2019; Tan & Bansal, 2019). Each example is two photographs, one sentence, and a true/false label. The test set is the official development split of $6{,}982$ examples. A photos-only example has both photographs, no sentence, and the label. A sentence-only example has the sentence, no photographs, and the label.

Model. We used LXMERT-base (Tan & Bansal, 2019) which has 210M parameters. Finetuning backpropogated through the entire model. After encoding both photographs, the two outputs were concatenated into a two-layer classifier. We used BertAdam optimizer with learning rate 5×10^{-5}, 10% warmup and linear decay, weight decay 0.01, gradient clipping at 5, batch size 32, sentence length 20, label smoothing 0.1, and a fresh shuffle each epoch. One sixteenth of training data was heldout as validation.

Metric. We used test cross-entropy as our loss after fitting temperature on validation logits, as we encountered improved accuracy but worsened cross-entropy in initial runs due to overconfidence.

C Robustness Checks

We test several factors for robustness of our claims in the main paper. We report accuracy instead of cross-entropy (C.1), vary architecture and model size (C.2), and change how we count paired data (C.3).

C.1 Accuracy Instead of Cross-Entropy

Figure 8 repeats Figure 2 in accuracy instead of cross-entropy, and Figure 9 repeats Figure 3 in per-position accuracy in Digits. Both metrics yield similar findings.

- **Synergy is gated.** In Figure 8, we observe delayed improved test accuracy for Digits and NLVR2, with no such gate for VQA excepting the $p = 0$ jump. This is the same as in cross-entropy.
- **VQA displays a $p = 0$ jump.** In VQA, accuracy rises from 0.37 to 0.42 with the first 4k pairs, then rises steadily to 0.49 at $p = 1$.
- **Presence-matching has a smaller effect for accuracy.** Unlike test cross-entropy, pairing has minimal effect on unimodal and redundant accuracy (Figure 9); at this data budget $D_1 = D_2 = 128$k, the U_1, U_2, and R slots are accurate up to the label noise threshold of 90%.

C.2 Architecture and Model Size

Digits. We replace the CNN encoders from Digits with two other encoders of different architectures, both trained from scratch on the same data: an MLP and a transformer. The MLP contains two layers of width 256 for each input digit and the transformer has sixteen patches per digit position with width 128, four layers, and four heads.

Figure 10 shows loss curves at $D_1 = D_2 = 32$k for the MLP and transformer architecture that are consistent with the CNN encoder. Test CE drops before synergy is learned, and S digit accuracy stays at chance then rises after a threshold. We note that architecture does seem to change where the threshold n_p is, with S digit accuracy first leaving chance at 16k pairs for the MLP, at 24k pairs for the transformer, and at 32k for the CNN.

VQA-v2 encoders. We replace frozen CLIP features with frozen Qwen2.5-VL-3B and Qwen2.5-VL-7B features (2048 and 3584 dimensions) under the same width-1024 head, at $D_1 = D_2 = 128$k. The Qwen image vector is the mean-pooled output of the vision encoder. The Qwen text vector is the mean of the question's input token embeddings with no transformer layers applied. Figure 11 (a) shows the same shape for all three encoders of a steep drop after $p = 0$ then a steady and gradual fall until $p = 1$.

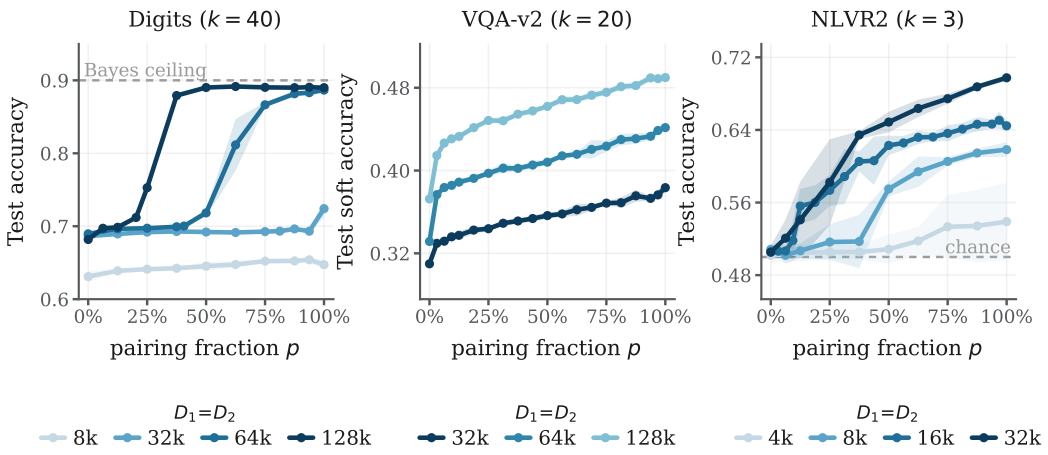


Figure 8: Figure 2 in accuracy instead of test cross-entropy.

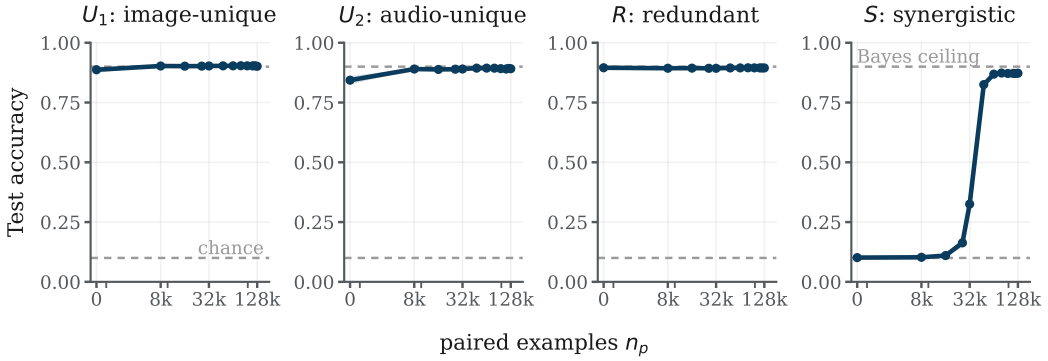


Figure 9: Figure 3 (Top) in accuracy instead of cross-entropy. Per-position test accuracy in Digits at $D_1 = D_2 = 128$k and $k = 40$ with both modalities shown.

VQA-v2 head width. Figures 11 (b) and (c) vary the MLP head width to 512 and 256. Each width observes the same behavior, and at fully paired data at $D_1 = D_2 = 128$k, all three heads end within 0.1 nats of each other. Model width changes the initial drop and gradual decline of the curve but not the curve shape.

C.3 Counting paired and unpaired data

The main text counts data per modality as $D_1 = n_1 + n_p$ and $D_2 = n_2 + n_p$. In this regime, a paired datum counts towards both D_1 and D_2. There are two other ways to count data:

- **Counting by label**, $D_1 + D_2 - n_p = n_1 + n_2 + n_p$, where a paired datum counts once.
- **Counting by price**, where a pair costs ϕ times as much as one unpaired example of each modality (Appendix A.5). If measuring by label and using the original definitions of D_1 and D_2, then a pair is actually cheaper than an unpaired datum with $\phi = \frac{1}{2}$.

Consequences of counting by label. Because the laws take (D_1, D_2, n_p) as input, the above decision of counting by label does not change any of the law fits, but it does change what "equal in size" means as well as deciding an optimal p under a budget. In Figure 12, we replot Figure 2 with number of labeled examples as the x-axis instead of paired examples. We find in this case, pairing is even more beneficial, as in addition to the synergistic gains demonstrated in this work, paired data also reduces the number of overall labels. This way of counting makes pairs cheaper and biases the optimal pairing ratio under a fixed budget towards more pairs.

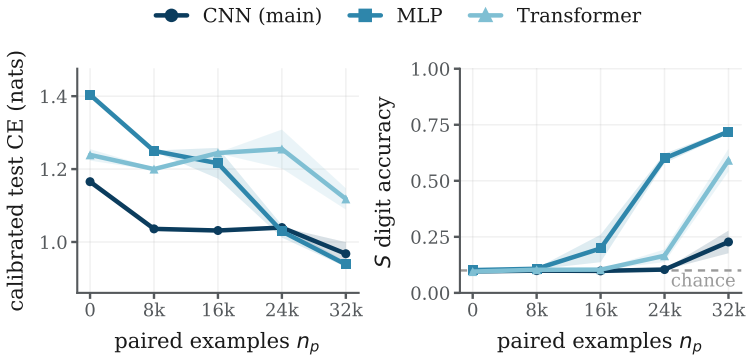


Figure 10: Digits CE (left) and S digit accuracy (right) versus n_p with three different encoders at $D_1 = D_2 = 32$k and $k = 40$.

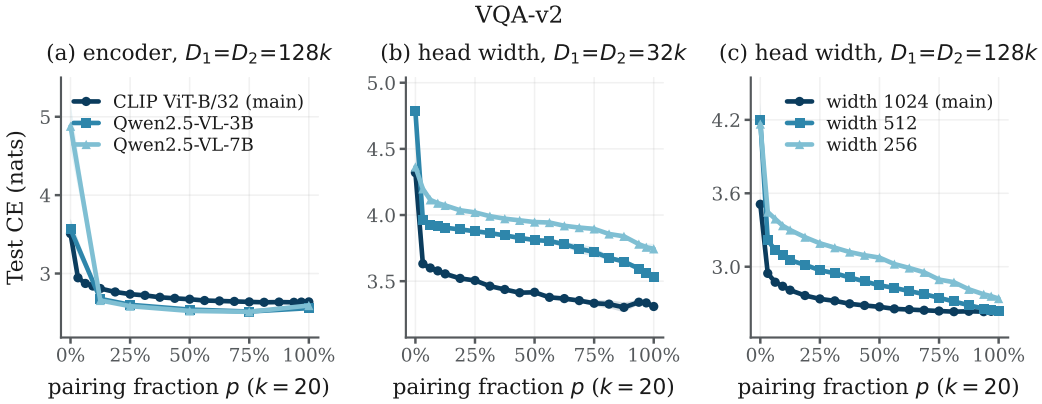


Figure 11: VQA-v2 test cross-entropy against p at $k = 20$. **(a)** Three frozen encoders under the width-1024 head at $D_1 = D_2 = 128$k. **(b, c)** Three head widths on CLIP features at 32k and 128k.

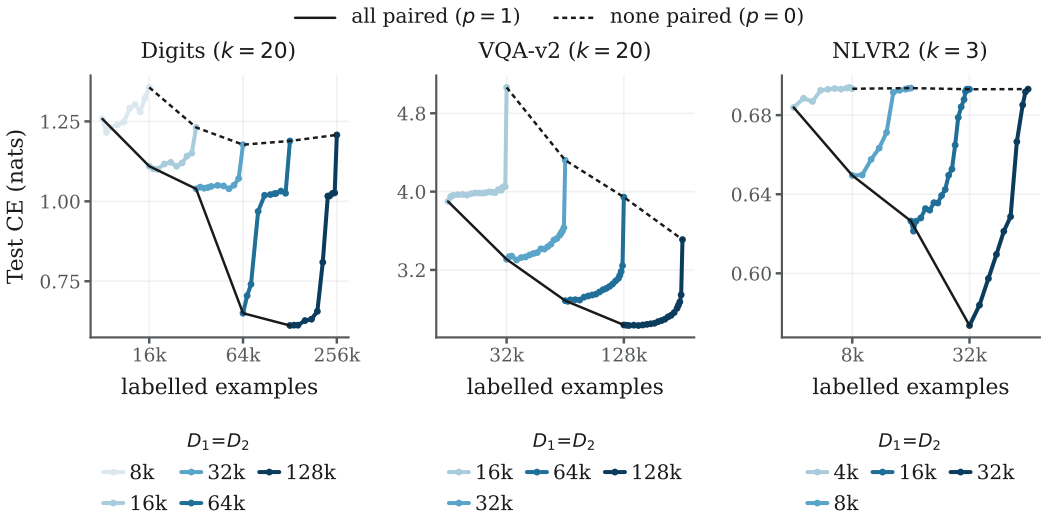


Figure 12: The runs of Figure 2 replacing the x-axis with number of labelled examples, $n_1 + n_2 + n_p = 2D - n_p$, instead of pairing fraction. Each colored curve represents one matched budget $D_1 = D_2 = D$ running from $p = 0$ (right point, $2D$ labeled examples) to $p = 1$ (left point, D labeled examples).