Scaling Laws for Generative Mixed-Modal Language Models

Armen Aghajanyan [*1] **Lili Yu** [*1] **Alexis Conneau** [1] **Wei-Ning Hsu** [1] **Karen Hambardzumyan** [2] **Susan Zhang** [1] **Stephen Roller** [1] **Naman Goyal** [1] **Omer Levy** [1] **Luke Zettlemoyer** [1 3]

Abstract

Generative language models define distributions over sequences of tokens that can represent essentially any combination of data modalities (e.g., any permutation of image tokens from VQ-VAEs, speech tokens from HuBERT, BPE tokens for language or code, and so on). To better understand the scaling properties of such mixed-modal models, we conducted over 250 experiments using seven different modalities and model sizes ranging from 8 million to 30 billion, trained on 5-100 billion tokens. We report new mixed-modal scaling laws that unify the contributions of individual modalities and the interactions between them. Specifically, we explicitly model the optimal synergy and competition due to data and model size as an additive term to previous unimodal scaling laws. We also find four empirical phenomena observed during the training, such as emergent coordinate-ascent style training that naturally alternates between modalities, guidelines for selecting critical hyper-parameters, and connections between mixed-modal competition and training stability. Finally, we test our scaling law by training a 30B speech-text model, which significantly outperforms the corresponding unimodal models. Overall, our research provides valuable insights into the design and training of mixed-modal generative models, an important new class of unified models that have unique distributional properties.

1. Introduction

Generative language models have been developed for a wide range of data modalities, including natural language text (Brown et al., 2020), code (Chen et al., 2021; Fried et al., 2022), images (Ramesh et al., 2021; Yasunaga et al., 2022), and molecules or proteins (Chilingaryan et al., 2022; Hsu et al., 2022). Recent work has also introduced unified models (Aghajanyan et al., 2022; Reed et al., 2022; Wang et al., 2022; Zellers et al., 2022) that can simultaneously model multiple modalities. One advantage of generative modeling in these cases is that the models scale well in practice; adding data, compute, or parameters typically improves model quality. These scaling trends have been carefully studied for uni-modal models (Kaplan et al., 2020; Hoffmann et al., 2022) and some recent work focuses on pairs of modalities (Droppo & Elibol, 2021; Henighan et al., 2020). However, the scaling behavior of larger number of modalities remains largely unstudied.

We present an extensive empirical study of scaling laws for mixed-modal generative language models over tokens. We assume that every modality can be represented as a sequence of tokens (e.g. VQ-VAEs for images (Esser et al., 2020) or HuBERT for speech (Hsu et al., 2021)). With this assumption, we can train a single discrete language model to represent data with arbitrary subsets of modalities presented in arbitrary orders. Such mixed-modal models are very general, but it is an open question the extent to which scale alone will be enough to overcome the inherent competition that comes as we add more modalities to a single model.

Through extensive experimentation, including over 250 individual experiments with seven modalities and model sizes ranging from 8 million to 30 billion, we have identified a scaling law that reflects the contributions of individual modalities and an additional term that captures the interaction between modalities (whether it be one of competition or synergy). We develop mixed-modal scaling laws that directly model competition between modalities and correctly predict data and model regimes where competition between modalities during training progresses into synergy. Specifically, we showed that our scaling laws correctly predicted the compute regime (30B model size, 45B token size), where we saw the complete reduction of modality competition for the Speech and Text modalities.

We also report a number of new empirical phenomena that arise during the training of mixed-modal models, including the tendency for the models to prioritize the optimization

[*]Equal contribution [1]FAIR [2]YerevaNN [3]University of Washington. Correspondence to: Armen Aghajanyan <armenag@meta.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

of a single modality at different stages of training. Our findings demonstrate that these phenomena can be primarily explained through the scaling law of interaction within the mixed-modal model. Additionally, we present new insights and guidelines for how to set key hyperparameters based on the terms of our scaling laws when optimal uni-modal hyper-parameters are known.

Our contributions are the following:

- We develop neural scaling laws for mixed-modalities models that include text, speech, images, code, and their numerous couplings.
- We discover a set of scaling laws describing the competition between arbitrary modalities.
- We provide a simple recipe for selecting hyper-parameters in a multi-modal setting when optimal uni-modal hyper-parameters are known.
- We uncover correlations between the scaling laws parameters we propose and various training phenomena, including training stability, optimal batch size, and coordinate ascent-like behavior in the optimization process across different modalities.

2. Related Work

Neural scaling laws quantify the relationship between model size, dataset size, compute budget, and performance, when training neural networks. This concept was introduced by Hestness et al. (2017), who observed a power law relationship and later scaled to much larger models by Kaplan et al. (2020).

Hoffmann et al. (2022) developed a unified formula for scaling laws, and provided recipes for compute-optimal training by adding data-dependent scaling terms unlike previous power law parameterizations. Other researchers have applied these principles to specific tasks and different parameterization of Transformers. Clark et al. (2022) examined the application of neural scaling laws to Mixture of Experts (MoE) models. Dettmers et al. (2022); Dettmers & Zettlemoyer (2022) studied the relationship between scaling laws and lower precision, which refers to using lower-precision data types, such as 16-bit floating point numbers, in neural networks. Gordon et al. (2021) and Ghorbani et al. (2021) applied these principles to Neural Machine Translation (NMT).

Additionally, Henighan et al. (2020) and Droppo & Elibol (2021) examined the application of neural scaling laws to generative language models in different modalities, including image generation and acoustic models. Cherti et al. (2022) also examined multi-modal training but did not specifically focus on generative models. To our knowledge, we are the first to investigate the phenomenon of interactions, competition, and interference between multiple modalities during training and provide a recipe for optimal mixed-modal training.

Interestingly, similar competition and scaling phenomenon have been observed for multi-lingual models. Conneau et al. (2019) observed a "curse of multilinguality," where training in multiple languages can lead to interference between languages, resulting in decreased performance. Goyal et al. (2021) and Shaham et al. (2022) demonstrated that this interference could occur even on models much smaller than the available training data, but scaling up the model size can improve synergy and alleviate interference. These findings align with our findings in the mixed-modal scenario, suggesting that similar principles apply when training on multiple modalities.

3. Definitions

3.1. What is a Modality?

Modalities are traditionally distinguished by the data source, domain, or sensor affinity. For example, the code domain is typically seen as distinct from text due to the different data involved (e.g., GitHub vs. CommonCrawl). This also applies to auditory or visual modalities, which are captured with different sensors. Yet the decisions are not always clear, for example different languages are often all within the domain of the text. Given that we are studying neural scaling laws across modalities, we aim to have an empirically testable modality definition.

We define $\sigma-$membership of a set of samples, D_α through the following membership function using perplexity as $\mathcal{L}$.

$$D_i \in D_j \iff \mathop{\mathbb{E}}_{x\sim D_i}\left[\mathcal{L}_{D_j}(x)\right] \leq \sigma^2 \mathop{\mathbb{E}}_{x\sim D_j}\left[\mathcal{L}_{D_j}(x)\right] \quad (1)$$

We empirically define **modality** by comparing the perplexity of one data set to another. Suppose the perplexity of the secondary data set over the probability distribution of the primary set is greater than σ times the mean perplexity of the primary set. In that case, we consider them to be distinct modalities. This definition distinguishes modalities by source, domain, sensor affinity, and language. We use the standard definition of perplexity (ppl). Using this definition with $\sigma = 3$, we decided to select seven modalities that we describe in detail below: `Text`, `Image`, `Image-Text`, `Speech`, `Speech-Text`, `Code`, `Molecules`.

Additionally, we define **source modality** as the type of token the sample contains, which within our setting will be; `Text`, `Speech`, or `Image`.

Functional Approximation Error
# of Model Parameters
For Modality j

$$\mathcal{L}\left(N, D_j\right) = E_j + \frac{A_j}{N^{\alpha_j}} + \frac{B_j}{|D_j|^{\beta_j}} \quad (2)$$

Dataset
Minimal Achievable Loss
Convergence Error

3.2. Uni-Modal Scaling Laws

We selected the Hoffmann et al. (2022) parameterization of scaling laws due to its precise representation of data factors and its additive nature, which allows for easy extension to multiple modalities. This parameterization (Equation 2) describes the loss based on the number of model parameters (N) and the number of tokens ($|D|$) through three constituent parts: the minimal achievable loss (E), the functional approximation error ($\frac{A_j}{N^{\alpha_j}}$), and the optimization or convergence error ($\frac{B_j}{|D_j|^{\beta_j}}$). These three factors are captured through seven learned parameters, providing a precise description of the loss.

It is well established that the upper bounds for β and α are both $\frac{1}{2}$, which provides a clear understanding of how well transformers coupled with gradient descent algorithms scale in relation to the optimal scaling for each modality (Hoffmann et al., 2022).

4. Empirical Setting

4.1. Datasets

Text For our text corpus, we use the same data as was used in OPT Zhang et al. (2022) for a total of 180B tokens. This dataset is primarily in English, although it contains other languages, as no explicit language filtering was done.

Image For all images, we convert them to discrete tokens using the Make-A-Scene visual tokenizer (Gafni et al., 2022), which gives 1024 tokens from an 8192 vocabulary per image. We select a custom subset of 600 million images across Schuhmann et al. (2022), and a custom image-text dataset scraped from Common Crawl. We remove all NSFW images and images that contain watermarks. Our `Image` dataset only contains the image and not the caption for a total of 614 billion tokens.

Image-Text We utilize the `Image` dataset described above but align it with captions available from the image for a total of 690 Billion tokens. We call this our `Image-Text` dataset.

Speech We used a combination of custom web-mined speech data and unlabeled speech in several public datasets. The web-mined speech dataset contains only unlabeled data in the form of long podcasts or news. We follow a series of preprocessing steps to improve the data quality and remove music and sensitive speech data. We also use a LangID model to select English-only speech. Our public data collection covers various speech styles and content topics, including LibriSpeech (Read-Books), CommonVoice in Read-Wiki, VoxPopuli from the Parliament domain, and Spotify Podcast and People's Speech as web speech. Thanks to this combination, our `Speech` dataset offers a rich diversity.

Speech-Text Many public datasets also come with text aligned with speech. We take ASR and TTS data from Multilingual LibraSpeech and VoxPopuli and form the `Speech-Text` dataset.

Code We use the InCoder data (Fried et al., 2022).

Molecules We utilize the Simplified Molecular Input Line Entry System (SMILES, where the chemical's structure is serialized into a string of symbols) representation from the Zinc dataset prepared by Chilingaryan et al. (2022).

4.2. Tokenization

Our mixed-modal generative models use a unified tokenization over all the modalities mentioned. This tokenizer processes data from all modalities into discrete tokens, which can be processed jointly by our model and trained with a single loss.

We use a Vector Quantized Variational autoencoders (VQGAN (Esser et al., 2020)) model to tokenize image data into discrete tokens. The VQGAN model compresses each image into a grid of image tokens, where an encoder encodes each token into a vector. This process reduces the context size of the transformer by a factor of $3 * X^2$, where $X$is the spatial reduction rate, or patch size, and 3 is the number of image channels. Online clustering is then performed, mapping each vector to the nearest entry of a learned codebook. We use a variant of the VQGAN from Gafni et al. (2022), which has a spatial reduction of 8 and a codebook size of 8192. This model is trained with extra perceptual losses to specific image regions, such as faces and salient objects, which improves the fidelity of the generated images. To be most effective in the language model stage, the visual tokenizer needs to effectively represent an image, and the correlated decoder needs to reconstruct the generated image tokens into high quality image data. We benchmark various image pretokenizers for those properties in Appendix 8.3.1.

We use a Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) model for tokenizing our speech data. HuBERT is a self-

supervised learning (SSL) model. It is trained to predict a masked subset of the speech signal using a mask language model objective, and has been found to be effective in learning a combined acoustic and language model over the continuous speech inputs. An offline clustering step to then used to generate discrete units. We use the BASE HuBERT model in our work (model and training details see the appendix 8.3.2). The final HuBERT units are generated through K-means clustering of the third iteration feature at the last layer, with a codebook size of 2000. Our HuBERT model produces embeddings at 50 Hz for input waveforms at 16 kHz, which is a 120-fold reduction in the sampling rate, while effectively retaining speech information (Analysis see 8.3.2).

Finally, we randomly sample 10 million sentences from all the data sets mentioned above and train a BPE model, where image and speech tokens take up a single token and a sentence represents a full image (1024 tokens) or an audio sample. We do an additional digit splitting for a vocab size of 2^{16} (Sennrich et al., 2016).

Given the high cost of running these experiments, we do not extensively explore the impact of different tokenization methods or vocabularies on the scaling laws. However, our findings provide a solid foundation for further investigating these factors.

4.3. Model Architecture

We study the family of decoder-only models described in GPT-3 Brown et al. (2020) and OPT Zhang et al. (2022). We limit ourselves to training up to 6.7 billion-parameter models for all our uni-modal and bi-modal scaling laws and train up to 30B parameters to measure the generalizability of our scaling laws. For completeness, we present model architecture and their respective sizes in Table 1. We use learned positional encodings across all model architectures.

4.4. Causal Masking Objective

Instead of the traditional left-to-right causal language modeling objective, we use the causal masked objective from Aghajanyan et al. (2022). This provides a form of bidirectional context for sequence infilling, and also supports more aggressive generalization. For example, causally masked models trained only on data with text followed by images can still flip the ordering to generate images from text, since they were not strictly trained to predict tokens left to right. Recent work also shows that this masking does not hurt language modeling performance or the generative capacity of the models (Fried et al., 2022; Bavarian et al., 2022). We provide additional support for this claim in § 8.2.

4.5. Training Procedure

All models were trained using the metaseq[1] code base, which includes an implementation of causal masking (Zhang et al., 2022). The training used the PyTorch framework (Paszke et al., 2019), with fairscale to improve memory efficiency through fully sharded model and optimizer states (Baines et al., 2021). The training also uses Megatron-LM Tensor Parallelism Shoeybi et al. (2019) to support large model runs, and we use `bf16` (Kalamkar et al., 2019) to improve training stability. Given the large volume of data, we performed a single epoch of training, using each training document once. The batch size per GPU was determined based on the total world size of the experiment, the level of model parallelism, and the total target batch size in terms of the number of tokens. To ensure stable training, we applied gradient clipping with a maximum norm of 1.0 and used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ (Kingma & Ba, 2015). We used the built-in polynomial decay learning rate scheduler in MetaSeq with 500 warmup updates and the end learning rate set to 10% of the peak learning rate.

We tracked all experiments using the Aim experiment tracker (Arakelyan et al., 2020). To ensure consistent training strategies across our experiments, we implemented a model restart policy using the Aim experiment tracker and callbacks. Specifically, if training perplexities do not decrease after 500 million tokens, the training run is restarted with a reduced learning rate with a factor of 0.8 of the learning rate from the last time step. This policy helps remove variance in the scaling laws due to differences in training procedures and allows us to scale up the number of asynchronous experiments significantly.

All experiments were conducted in a two-month time frame with a cluster of 768 80GB A100 GPUs. The majority of experiments used 64 GPUs at a time.

5. Scaling Laws

5.1. Uni-Modal Scaling Laws

We first aim to discover scaling laws for each of the individual modalities we listed above. We train seven different model sizes, from 8 million to 6.7 billion, on seven different modalities on three different dataset sizes (5B, 10B, 100B).

In Figure 12 available in the appen, we share the training curves for all modalities and model sizes for the largest data size (100B tokens), and the final performance of all models in Figure 1. Overall, we see that scaling dynamics are fundamentally different across modalities, scale, and dataset size (which further reinforces our selection of dataset-size-dependent parameterization of scaling laws).

[1] https://github.com/facebookresearch/metaseq

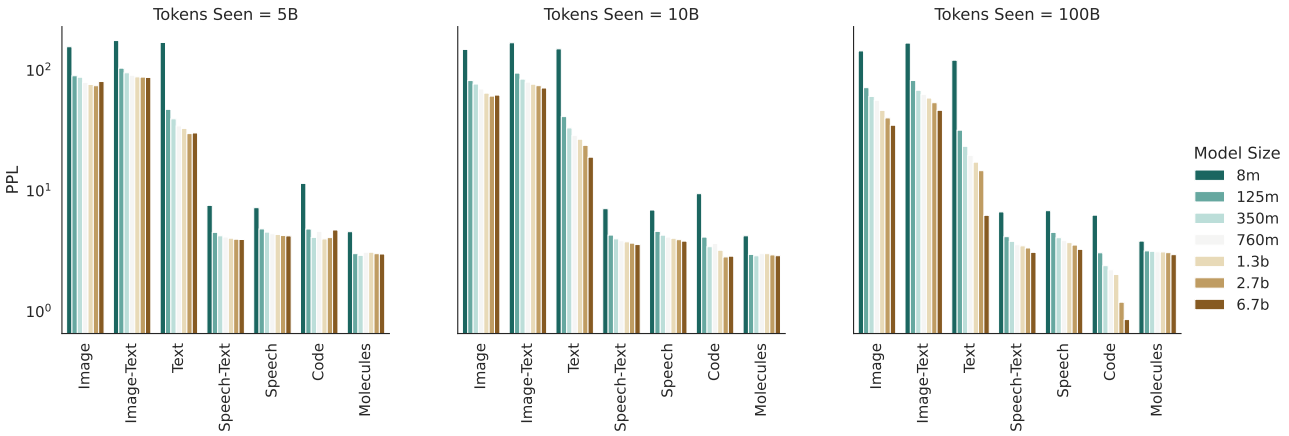

Figure 1. Empirical scaling properties across both data and model size scale for the uni-modal setting.

For each modality, we fit the seven parameters from Equation 2, following the procedure in Hoffmann et al. (2022). Specifically, we minimize

$$\sum_{\substack{\text{run } i \\ \text{mod } j}} \underset{\sigma=0.03}{\mathrm{H}} \left[\mathrm{LSE}(a_j - \alpha_j \log N_i, b - \beta \log D_i, e_j) - L_i\right] \tag{3}$$

for $\{a_j, b_j, e_j, \alpha_j, \beta_j\}$ where H is the standard Huber loss, for every run i and modality j.

We then set $A_j = e^{a_j}$, $B_j = e^{b_j}$, $E_j = e^{e_j}$. In order to identify the optimal minima, we followed the method outlined by Hoffmann et al. (2022) and employed the L-BGFS algorithm on the same grid of initialization values. Our only deviation was using a higher value for the Huber loss parameter σ, which was necessary for generalization to held-out data in our multi-modal setting. The optimal values obtained were not located on the boundaries of the initialization grid.

The scaling laws for each modality are presented in Table 5 available in the Appendix. The parameters for each modality vary significantly. Some modalities, such as `Code` and `Molecules`, demonstrate more efficient use of the power of scale compared to others, such as `Image`. Our coefficients for `Text` are similar to those reported by Chinchilla, although it should be noted that we used a different dataset for our analysis. This accounts for any differences in the results.

5.2. Bi-Modal Scaling Laws

We also estimate scaling laws for training on two modalities: $\mathcal{L}(N, D_i, D_j)$, where N represents the model size, and D_i and D_j represent the two datasets being used. In the case where D_i and D_j are completely independent and have no mutual information between them, we expect the minimal achievable loss to be the average of the two monomodal scaling laws, given by $0.5\left[\mathcal{L}(\infty, D_i) + \mathcal{L}(\infty, D_j)\right]$. This is because we are averaging over the loss and subsampling both test datasets equally ($|D_i| = |D_j|$). On the other hand, if there is some form of mutual information present between D_i and D_j, we can expect the loss to be reduced by some maximal factor $C_{i,j}$. When considering finite model size and data regimes, there will be competition between the function approximation and optimization processes, which can be modeled using the same form as in Equation 2. We present our scaling law for mixed modal models in Equation 4.

$$\mathcal{L}(N, D_i, D_j) = \underbrace{\left[\frac{\mathcal{L}(N, D_i) + \mathcal{L}(N, D_j)}{2}\right]}_{\text{Loss if Datasets Were Modeled Independently}} - \underbrace{\mathcal{C}_{i,j}}_{\text{Max Level of Synergy}} + \underbrace{\frac{A_{i,j}}{N^{\alpha_{i,j}}}}_{\text{Competition in Functional Approximation}} + \underbrace{\frac{B_{i,j}}{|D_i| + |D_j|^{\beta_{i,j}}}}_{\text{Competition in Optimization Process}} \tag{4}$$

An additional benefit to this parameterization is the additive or linear nature, which allows us to extend our parameterization to n-modal scaling laws.

We justify this functional form by viewing it as an extension of the Chinchilla scaling laws, that directly models the synergy or competition between modalities in the same functional form as Chinchilla.

5.2.1. Experimental Results

We selected seven different pairs: `Image-Text|Code`, `Image-Text|Speech-Text`, `Image-Text|Text`, `Speech|Text`, `Code|Text`, `Molecule|Code`, and `Speech|Code`. While other couplings are available, we cannot do an exhaustive sweep due to computational constraints. We selected these pairs to maximize variety. For example, while `Code|Text` is known to perform well, `Image-Text|Code` may not offer as much benefit.

We create a dataset for each coupling and token target count where each subdataset contributes 50% of the tokens. We train using the same hyper-parameters as the uni-modal trainings and fit the scaling laws per modality coupling using the same procedure and optimization process (Equation 3). We present the empirical results in Figure 2.

5.2.2. Breaking The Competition Barrier

Given these laws, we can now make predictions about what scale will be required to overcome modal competition and achieve synergy from training on each pair of modalities. By modality competition, we refer to the empirical phenomena of two modalities performing worse than if we trained two individual models on the same number of per-modality tokens. By synergy, we mean the inverse. We can define the notion of synergy formally through our scaling laws. If

$$\mathcal{C}_{i,j} > \frac{A_{i,j}}{N^{\alpha_{i,j}}} + \frac{B_{i,j}}{|D_i| + |D_j|^{\beta_{i,j}}} \tag{5}$$

we are reducing the loss beyond the independent modeling of the modalities and therefore are synergistic; otherwise, we say the modalities are in competition. When both sides of the inequality are equal, we call this the competition barrier for the two modalities. We present our extrapolated scaling laws with the predicted competition barrier in Figure 3.

We can then find the compute-optimal model size and token count that breaks the competition barrier by minimizing a compute cost over the competition barrier. We select the approximation from Kaplan et al. (2020).

$$\begin{aligned} &\min_{N,|D|} 6ND \\ &\text{s.t. } \mathcal{C}_{i,j} = \frac{A_{i,j}}{N^{\alpha_{i,j}}} + \frac{B_{i,j}}{|D_i| + |D_j|^{\beta_{i,j}}} \end{aligned} \tag{6}$$

For the `Speech|Text` coupling, the predicted compute optimal parameters are $N = 28.35$B and $D = 45.12$B. To test this hypothesis, we select the closest architecture available from Zhang et al. (2022), which is the 30B parameterization and 50B tokens, slightly above the predicted data regime, to cover any error in our approximation. We train three models a, 350M, 2.7B, and 30B models on either `Speech`, `Text`, or `Speech|Text`. We plot the ratio of the average of the `Speech` and `Text` models perplexity per timestep by `Speech|Text` perplexity, the competition barrier and predictions from our scaling laws in Figure 4. As we see, the prediction does hold, and we achieve a model that crosses the competition barrier. Further scaling is likely to further improve the synergy, but we leave this exploration to future work.

6. Emergent Phenomena

We observed a number of emergent behaviors during training, many of which can be predicted from the modality-specific constants in our scaling laws. We briefly document these behaviors here; each is potentially worthy of study in future work.

Phenomenon 1. ***Intermittent Coordinate Ascent Like Training:*** *Different source modalities in a multi-modal setting are optimized at different paces, with some modalities even pausing their training progression for a significant amount of steps.*

When looking at average perplexity over the dataset, the training dynamics are always consistently smooth and somewhat monotonically decreasing (Figure 12). But looking at the sub-perplexities of the modalities shows a different picture; certain modalities flatten out during training (see left figure in Figure 5). In Figure 6, we plot the percent of the submodality that exhibits flatness, where flatness is defined as an area of the training curves where loss does not decrease (we do not count the warm-up period of optimization as part of this percentage).

Phenomenon 2. ***Rate of Phenomena 1 Diminishes Past A Certain Scale:*** *The rate of intermittent coordinate ascent-like training is correlated with scale (N) and $\alpha_{i,j}$.*

Most of this intermittent coordinate ascent-like training can be reduced by simply increasing the model size. Intuitively, this makes sense as the increased functional approximation space should give the models enough capacity to simultaneously optimize all of the modalities (Figure 6). Additionally, we discover that the empirically found $\alpha_{i,j}$, which describes the functional approximation cost across two modalities, is highly correlated with the uni-modal optimization flatness in the training regime. We found no correlation between $\beta_{i,j}$ and optimization flatness.

Phenomenon 3. ***Optimal Batch Size for Modalities i and j is Correlated with $\beta_{i,j}$***

We fixed the batch size to 1M tokens, but the question of the optimal batch for each modality and modality coupling remains. We train four versions of all over a subset of models, overall modalities, and selected couplings of modalities with batch sizes 1M, 2M, 4M, and 8M over 5B tokens, with the exception of modalities that contain `Text` for which we add 0.5M batch size experiments. We use the same training regime as mentioned in § 4.5. We present our results in Figure 11. Additionally for the bi-modal coupling experiments we plot $\log$ of the ratio between the optimal batch size and the sum of the optimal batch sizes for the sub-datasets against the $\beta_{i,j}$ of the discovered scaling laws in § 5.2. We found no correlation between $\alpha_{i,j}$ and optimal batch-size.

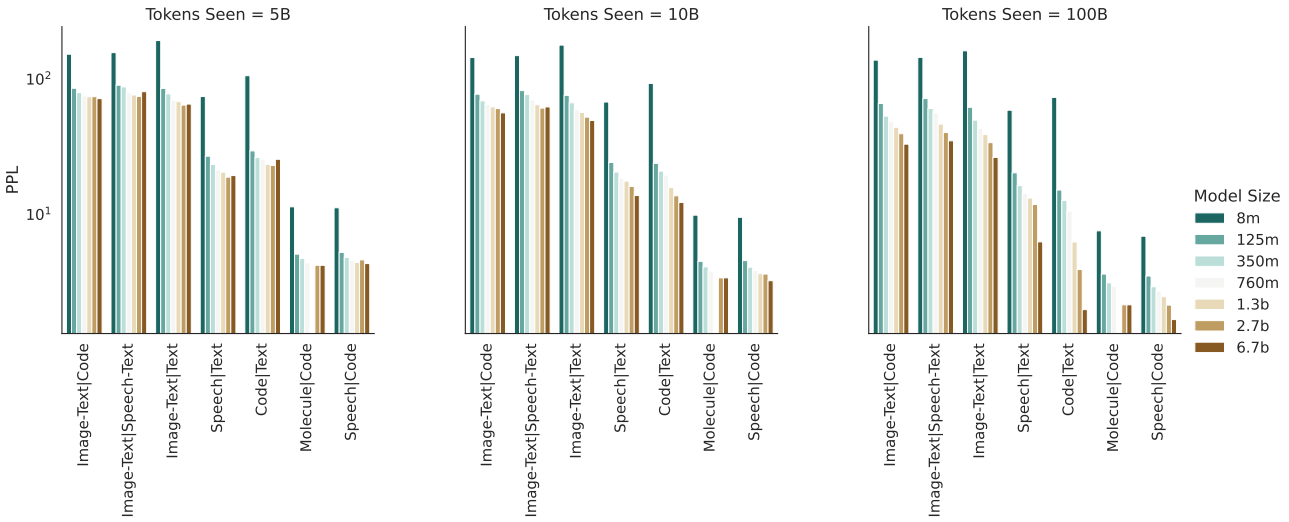

Figure 2. Empirical scaling properties across both data and model size scale for the multi-modal setting.

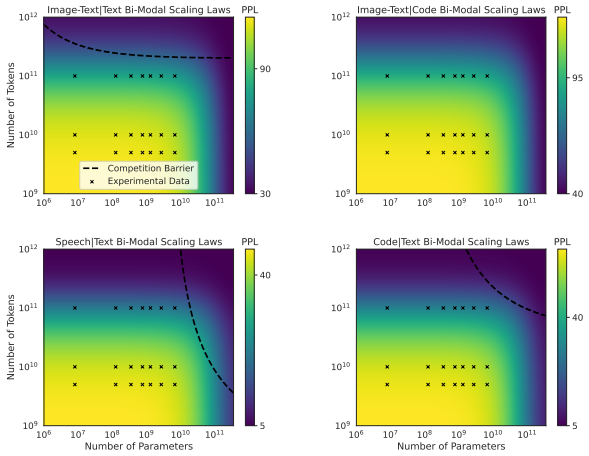

Figure 3. Bi-modal scaling law extrapolation and the predicted competition barrier for a subset of our bi-modal experiments.

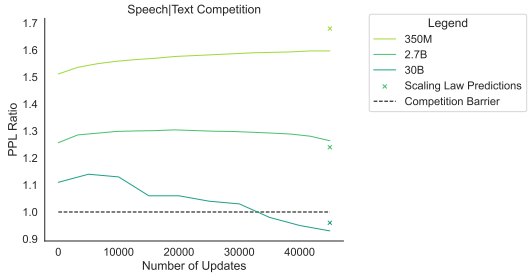

Figure 4. We plot $\frac{\mathcal{L}(N,[\texttt{Speech},\texttt{Text}])}{0.5*(\mathcal{L}(N,\texttt{Text})+\mathcal{L}(N,\texttt{Speech}))}$ throughout the training process. If this ratio is below 1, we have broken through the competition barrier. Additionally, we add the predictions for the final ratio as predicted from our scaling laws.

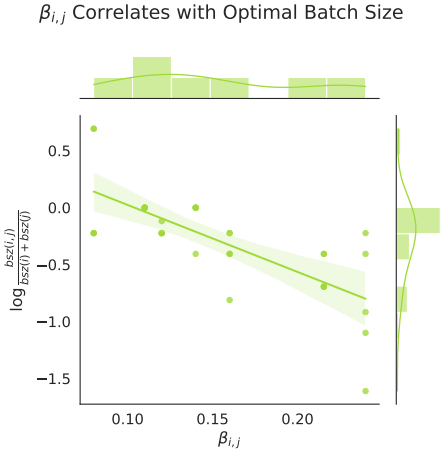

Figure 7. The logarithm of the ratio between the optimal batch size for the entire dataset and the sum of the optimal batch sizes for the individual sub-datasets plotted against the $\beta_{i,j}$ values of the scaling laws that were identified in the previous section.

Phenomenon 4. ***Rate of Deteriorating Training Dynamics is Correlated with $\alpha_{i,j}$ and N***

The stability of training can be captured by looking at the total count of gradient norm spikes throughout the lifetime of the training. A large number of gradient spikes can indicate a poor training setting, from selecting the wrong learning rate or batch size to having low-quality data. Additionally, larger models tend to be harder to stabilize, reflecting in a larger amount of gradient spikes. We hypothesize that lower values of $\alpha_{i,j}$, reflecting higher competition between modalities, will correlate with more gradient norm spikes. We present the empirical correlation between $\log(N)/\alpha_{i,j}$ and # of Gradient Norm Spikes in Figure 8. We see a highly predictive relationship between model size (N) and the rate

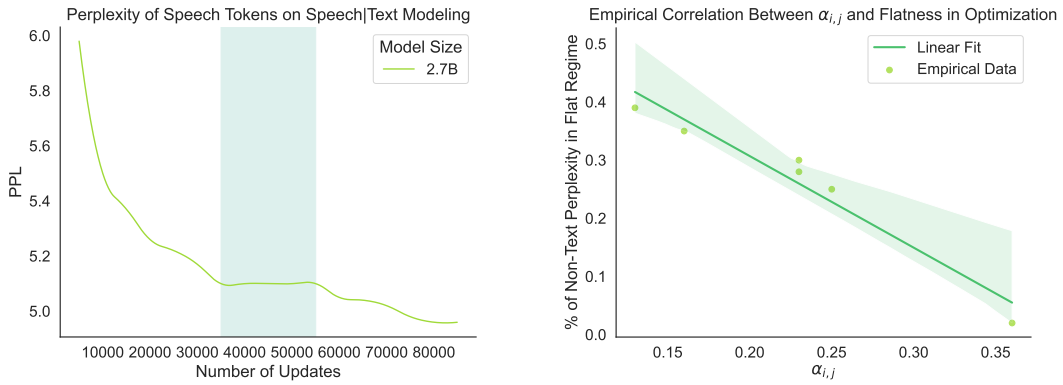

Figure 5. **Left:** Example run showing the perplexity on only the speech tokens of a 2.7B run over the `Speech|Text` dataset. We highlight a region where roughly for 15000 steps perplexity for speech flattened. **Right:** Correlation between the mixed-modal $\alpha_{i,j}$ parameter and the percent of non-text perplexity that are within a flat regime in the 6.7B model regime.

Figure 6. Percent of the submodality that exhibits flatness, where flatness is defined as an area of the training curves where loss does not decrease. We present these plots for speech and image perplexity within the bi-modal couplings that contain them.

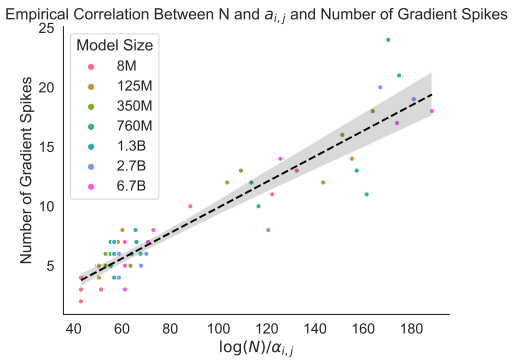

Figure 8. We plot $\log(N)/\alpha_{i,j}$ against the number of gradient spikes that occurred for the respective experiments.

of mixed-modal competition ($a_{i,j}$ to the stability of the training run. We found no correlation between $\beta i, j$ and the # of gradient norm spikes.

7. Conclusion

We have provided extensive experimentation and analysis into the scaling properties of mixed-modal generative models. By developing a scaling law that reflects the contributions of individual modalities and the interaction between them, we have gained a deeper understanding of scaling mixed-modal models and the training dynamics of these models. Our findings also include a set of empirical phenomena observed during the training process and training dynamics that can be primarily explained through various interaction terms in our newly proposed scaling law. Additionally, we have developed guidelines for selecting critical hyper-parameters based on our scaling law, providing a valuable tool for practitioners in the field. Overall, our research has advanced the knowledge and understanding of mixed-modal generative models and will help develop unified models that can handle multiple modalities simultaneously.

References

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Arakelyan, G., Soghomonyan, G., and The Aim team. Aim, 6 2020. URL https://github.com/aimhubio/aim.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Baines, M., Bhosale, S., Caggiano, V., Goyal, N., Goyal, S., Ott, M., Lefaudeux, B., Liptchinsky, V., Rabbat, M., Sheiffer, S., Sridhar, A., and Xu, M. FairScale: A general purpose modular PyTorch library for high performance and large scale training. https://github.com/facebookresearch/fairscale, 2021.

Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., and Chen, M. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

Chilingaryan, G., Tamoyan, H., Tevosyan, A., Babayan, N., Khondkaryan, L., Hambardzumyan, K., Navoyan, Z., Khachatrian, H., and Aghajanyan, A. Bartsmiles: Generative masked language models for molecular representations. *arXiv preprint arXiv:2211.16349*, 2022.

Clark, A., de Las Casas, D., Guy, A., Mensch, A., Paganini, M., Hoffmann, J., Damoc, B., Hechtman, B., Cai, T., Borgeaud, S., et al. Unified scaling laws for routed language models. In *International Conference on Machine Learning*, pp. 4057–4086. PMLR, 2022.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Dettmers, T. and Zettlemoyer, L. The case for 4-bit precision: k-bit inference scaling laws, 2022. URL https://arxiv.org/abs/2212.09720.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

Droppo, J. and Elibol, O. Scaling laws for acoustic models. In *Interspeech 2021*, 2021. URL https://www.amazon.science/publications/scaling-laws-for-acoustic-models.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis, 2020.

Fried, D., Aghajanyan, A., Lin, J., Wang, S., Wallace, E., Shi, F., Zhong, R., Yih, W.-t., Zettlemoyer, L., and Lewis, M. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*, 2022.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., and Cherry, C. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*, 2021.

Gordon, M. A., Duh, K., and Kaplan, J. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.478. URL https://aclanthology.org/2021.emnlp-main.478.

Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M., Ali, M., Yang, Y., and Zhou, Y. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Kalamkar, D., Mudigere, D., Mellempudi, N., Das, D., Banerjee, K., Avancha, S., Vooturi, D. T., Jammalamadaka, N., Huang, J., Yuen, H., Yang, J., Park, J., Heinecke, A., Georganas, E., Srinivasan, S., Kundu, A., Smelyanskiy, M., Kaul, B., and Dubey, P. A study of bfloat16 for deep learning training, 2019. URL https://arxiv.org/abs/1905.12322.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Lee, A., Chen, P.-J., Wang, C., Gu, J., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., et al. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-maron, G., Giménez, M., Sulsky, Y., Kay, J., Springenberg, J. T., Eccles, T., Bruce, J., Razavi, A., Edwards, A., Heess, N., Chen, Y., Hadsell, R., Vinyals, O., Bordbar, M., and de Freitas, N. A generalist agent. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=1ikK0kHjvj. Featured Certification.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Shaham, U., Elbayad, M., Goswami, V., Levy, O., and Bhosale, S. Causes and cures for interference in multilingual translation. *arXiv preprint arXiv:2212.07530*, 2022.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2019. URL https://arxiv.org/abs/1909.08053.

Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, 2022. URL https://arxiv.org/abs/2202.03052.

Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*, 2022.

Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., and Choi, Y. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16375–16387, June 2022.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V.,

et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

8. Appendix

8.1. Model Architecture

All models are trained with pre-norm and using ReLU activation. We apply a dropout of 0.1 throughout, but we do not apply any dropout to embeddings. We also use weight decay of 0.1. To initialize the weights, we use a variant based on Megatron-LM codebase, which involves using a normal distribution with a mean of zero and a standard deviation of 0.006. We truncate this normal distribution within two standard deviations and observed substantial gain in both training stability and performance.

8.2. Causal Masked vs. Causal Objective

We measure the impact of the choice of objective by conducting an additional scaling law on our `Speech` and `Text` datasets on the standard (causal) language modeling objective. Everything is kept constant except for the objective, including the training procedures. We present the empirically fit scaling law parameters in Table 2.

Note that both objectives optimize the joint probability of tokens; therefore, if there was a significant difference in our perplexity, we should expect to see it reflected in a difference in scaling law parameters. Instead, we see that the scaling laws seem to be close to identical, with whatever minor differences within the error of our approximation.

8.3. Tokenization

8.3.1. Quality of Image Tokenization

Modeling long-range dependencies with raw pixel input of an image (for example, total sequence length for a 256-pixel image in RGB form is 196608) is non-trivial, especially with transformers, which in their vanilla form scale poorly with sequence length. Recently, Vector Quantized Variational autoencoders (VQ-VAE, or Discrete-VAE) have been proposed, which learn discrete image representations, allowing a later generative model to generate images in the discrete latent space, just like a standard language model. VQ-VAE reduces the context size of a transformer by a factor of $3 * X^2$ (X is the spatial reduction rate, and 3 is the number of image channels), where information loss is unavoidable. VQ-VAE is trained to optimize the evidence lower bound of distribution of data. (Esser et al., 2020) introduced VQGAN, which improves upon VQVAE by introducing an adversarial loss produced by a discriminator, reconstructing images with much higher quality. Recently, (Gafni et al., 2022) trained a new image tokenizer with a better training objective focusing on faces or objects, which is adopted for this work and denoted as $\text{VQGAN}_{\text{MAS}}$. To be most effective in the later language model stage, the image tokenizer must represent an image effectively. The corre-

Model	#L	dmodel	#H	dhead	Batch Size	LR	Context Length
8M	4	128	2	64	1M	1.00E-03	2048
125M	12	768	12	64	1M	6.00E-04	2048
350M	24	1024	16	64	1M	3.00E-04	2048
760M	24	1536	16	96	1M	2.50E-04	2048
1.3B	24	2048	32	64	1M	2.00E-04	2048
2.7B	32	2560	32	80	1M	1.60E-04	2048
6.7B	32	4096	32	128	1M	1.20E-04	2048
30B	48	7168	56	128	1M	1.00E-04	2048

Table 1. Model architecture details. We report the number of layers (#L), the embedding size (d_{model}), the number of attention heads (#H), the dimension of each attention head (d_{head}), batch size, learning rate (LR) and context length (# of tokens).

	A	B	E	α	β
Speech (CM3)	154.45	205.10	3.02	0.31	0.24
Speech (Causal)	164.12	201.00	3.01	0.30	0.24
Text (CM3)	492.51	1987.40	2.42	0.18	0.22
Text (Causal)	485.16	1859.32	2.45	0.17	0.23

Table 2. Uni-Modal scaling law parameters fit to Equation 2 for both causal (standard language modeling) and causal masked (CM3 objective from Aghajanyan et al. (2022)).

lated decoder must reconstruct the generated image tokens into high-quality image data. We benchmark the following pre-trained tokenizers on these properties:

- VQGAN(fx, y) with different spatial reduction rate fx and diffent vocab size y. For example, for a 256px image, 256 tokens will be created with a VQGAN$(f16)$ tokenizer and 1024 tokens with a VQGAN$(f8)$ tokenizer.
- Our VQGAN$_{\text{MAS}}256$ and VQGAN$_{\text{MAS}}512$ use $f8$ and $f16$ spatial reduction, respectively, and have an 8192 vocab size. Our VQGAN$_{\text{MAS}}256$ is trained with a face-aware loss with the help of a pre-trained face embedding model. Our VQGAN$_{\text{MAS}}512$ is trained with face+object aware loss with extra downsampling and upsampling layer in the encoder and decoder to reconstruct images with higher resolution. Note, Our VQGAN$_{\text{MAS}}512$ with 512x512 image input compresses the image to 1024 tokens, thanks to the downsampling layer.

One way to quantify the realism captured by these models is to compute Fréchet Inception Distance (FID) scores of reconstructed images w.r.t. the inputs (R-FIDs). Table 3 shows R-FIDs when reconstructing the whole validation split of the ImageNet dataset. For an image with a 256-pixel resolution, reducing spatial reduction rate or increasing visual vocab size can help achieve lower R-FIDs. Our VQGAN$_{\text{MAS}}256$ model is superior to its counterpart with the same spatial reduction rate and vocab size, demonstrating the effectiveness of the extra face-aware loss. Interestingly, VQGAN$_{\text{MAS}}256$ gets a higher R-FID than VQGAN$_{\text{MAS}}512$, consistent with the result in the original paper. We are also interested in understanding how much the reconstruction process can retain information and if we lose critical image information. We benchmark the representation power via classification accuracy with a pretrained model. We first reconstruct all the images in the ImageNet validation set with different tokenizers, similar to R-FID computation. Then a trained pretrained classifier on ImageNet is used to run inference on the original and reconstructed images. The classification accuracy of original images with 256 or 512-pixel resolution is 81.56 and 82.89, respectively. The accuracy@1 on the reconstructed images and their gap with the raw images are reported in Table 3. Images reconstructed by the VQGAN$_{\text{MAS}}256$ can best maintain the original information with less than 2 percent degradation in accuracy.

For qualitative comparison, we give all tokenizers except VQGAN$_{\text{MAS}}512$ an image with 256x256 pixels. VQGAN$_{\text{MAS}}512$ reconstructs a 512-pixel resolution image and resizes it to 256-pixel resolution for plotting purposes. VQGAN $(f16)$ a s produce 256 tokens, while VQGAN $(f8)$ and VQGAN$_{\text{MAS}}$ models produce 1024 tokens.

We randomly sample two images from ImageNet (top 2 rows in Figure 9). All reconstructed images can maintain vital information about the image and the textures. With a high reduction rate (192), VQGAN with $f16$ spatial reduction can not reproduce every detail of its input but tends to hallucinate parts of it, for example, the eye and the tail of the dog in row 1 and the mirror of the blue car in row 2. By increasing

Tokenizer	Spatial Reduction	Vocab	Token Counts	R-FID	Accuracy@1
256x256 px					
VQGAN	16	1024	256	7.94	69.25 (-12.4)
VQGAN	16	16384	256	4.98	73.2 (-8.45)
VQGAN	8	8192	1024	1.49	79.64 (-2.01)
VQGAN	8	16384	1024	1.14	79.47 (-2.18)
$\text{VQGAN}_{\text{MAS}}256$	8	8192	1024	**0.87**	**79.83 (-1.82)**
512x512 px					
$\text{VQGAN}_{\text{MAS}}512$	8	8192	1024	1.43	79.87 (-3.02)

Table 3. R-FID and accuracy with images reconstructed by a selection of image tokenizers

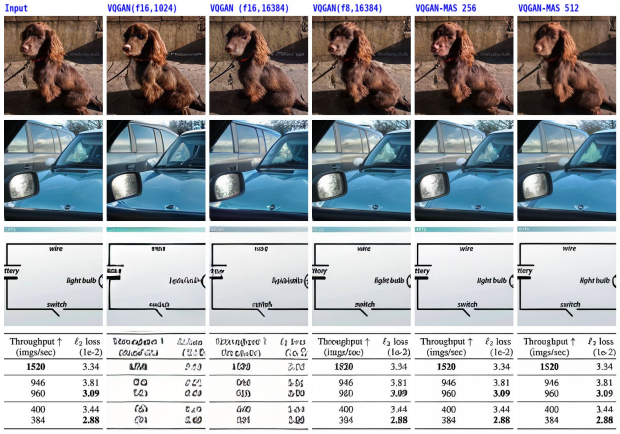

Figure 9. Reconstructed images across a selection of different image tokenizers.

the vocab size, more realistic images can be generated. With a decreased compression rate, the VQGAN $(f8)$ model and $\text{VQGAN}_{\text{MAS}}$ produce much more realistic reconstructed images. For example, in row 2, the door handle, the clouds, and the mirror's tree are successfully reconstructed with great detail.

Lastly, we reconstruct images from a textbook (row 4 in Figure 9) or screenshots of tables from scientific papers (row 5 in Figure 9). All models struggle to reconstruct the original image, except $\text{VQGAN}_{\text{MAS}}$ models. Figure 9 shows impressive results by $\text{VQGAN}_{\text{MAS}}$ models that all text and numbers are human readable. $\text{VQGAN}_{\text{MAS}}256$ produces sharper edges while $\text{VQGAN}_{\text{MAS}}512$ smooths things out.

From all the above examples, the reduced spatial reduction is effective for better tokenization; however, it results in a longer token sequence. Another way to increase image representation is to increase the pixel numbers of images. We reconstruct images with a size of 512x512 for $\text{VQGAN}_{\text{MAS}}512$ in Figure 10. VQGAN$(f16)$ produce 1024 tokens, while VQGAN$(f8)$ and $\text{VQGAN}_{\text{MAS}}$ models produce 4096 tokens. $\text{VQGAN}_{\text{MAS}}256$ in Figure 9 outperform VQGAN (f16) by a big margin. With the same token budget, decreasing spatial reduction is more effective than increasing image pixels.

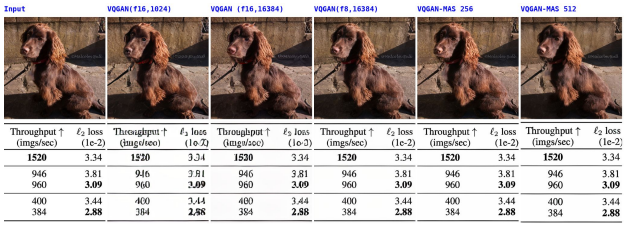

Figure 10. Reconstructed images across a selection of different image tokenizers (image size 512)

8.3.2. Details of Speech Tokenization

We use the Base HuBERT model in our work. This model comprises a convolutional encoder and 12 layer Transformer, each with an embedding dimension of 768, a feed-forward layer dimension of 3072, and 12 self-attention heads. Pre-training of the model has been performed on 32 GPUs over three iterations, with 400K updates per iteration. The training data consists of 221K hours of unlabeled speech from multilingual Librispeech (MLS) (Pratap et al., 2020), Common Voice (CV) (Ardila et al., 2019), and VoxPopuli (VP) (Wang et al., 2021) in eight languages (English, Spanish, French, German, Dutch, Italian, Polish, Portuguese). The MFCC/6-th layer feature from iteration 1 and the 9-th layer feature from iteration 2 are used as targets, with codebook sizes of 100/500/1000, respectively, following the methodology outlined in (Lee et al., 2021).

A typical 16kHz audio with a bit depth of 16 has a bitrate of 64kbps. HuBERT encodes audio at 50Hz with a codebook size of 2000, resulting in a bitrate of 548bps. The effective compression rate is roughly 117. Our model still

effectively retains speech information, as shown in Table 4. We compare the word error rate (WER) of a pretrained automatic speech recognition (ASR) model with original audio or reconstructed audio by HuBERT models. We present results with two HuBERT models, one public (Hsu et al., 2021) version (HuBERT public) and one trained by us (HuBERT ours). WER of the original audio on LJSpeech is 2.04, the audio reconstructed by HuBERT public degrades by 0.94, while the audio reconstructed by our HuBERT only degrades it by 0.3. A similar phenomenon is observed on the LibriSpeech dataset, where our HuBERT model improves upon HuBERT public and can effectively reconstruct audio with very little information loss.

8.4. Optimal Batch Sizes

We present the optimal batch sizes found in Figure 11.

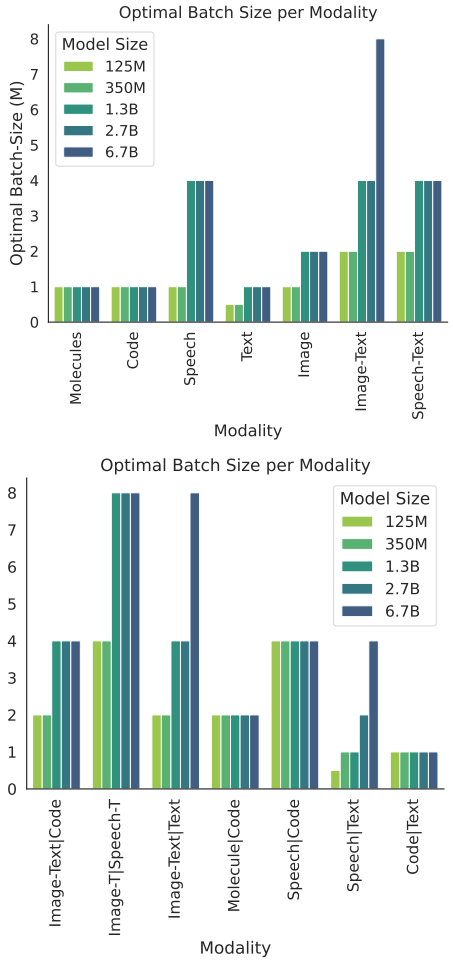

Figure 11. Optimal batch-size per modality and modality couplings across model sizes.

8.5. Training Dynamics in Uni-Modal Setting

Model	PT/KM data	Vocoder data	#L	K	LJSpeech	LibriSpeech
Orig audio					2.04	3.55
HuBERT public	LS960	LJ	9	500	2.98	12.39
HuBERT ours	MLS+VP+CV	LJ + MLS-40h	12	2000	2.34	9.06

Table 4. Word error rate (WER) on LJSpeech and LibriSpeech datasets of an petrained automatic speech recognition (ASR) model with various speech inputs (original audio, or recutructed audio by HuBERT model in (Hsu et al., 2021), or HuBERT model in our work). We also listed the model details of two HuBERT models, including data used during pretraining (PT), k-means (KM) and vocoder, number of layers (#L), and number of clusters (K).

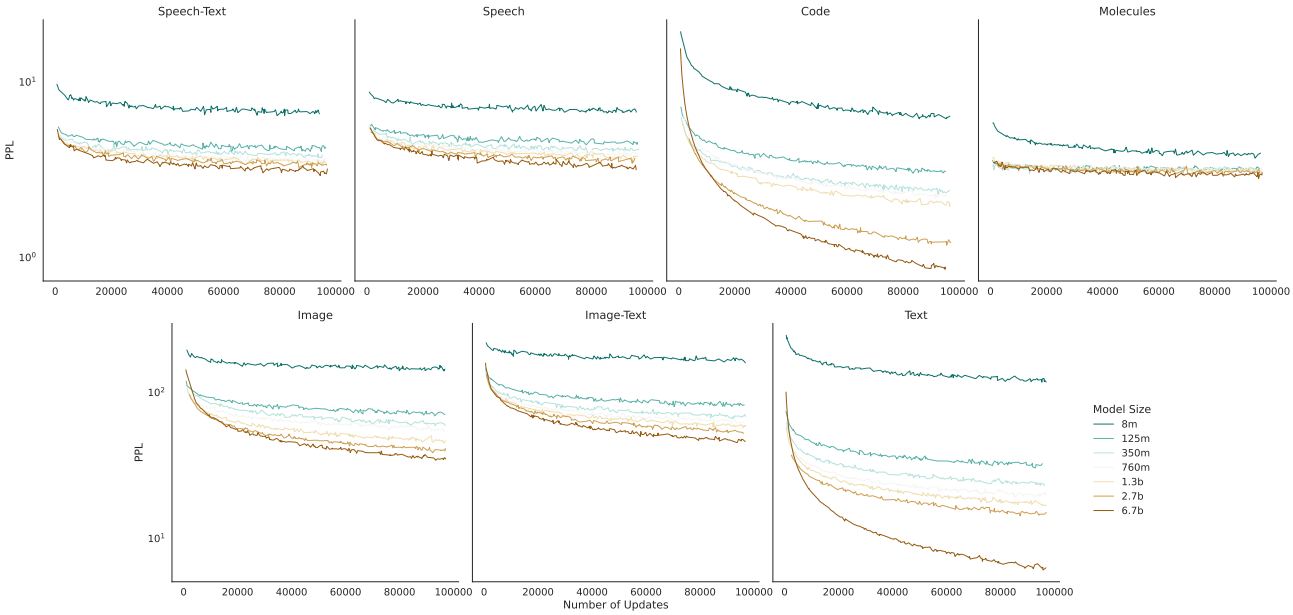

Figure 12. Single modality training curves for 100B tokens across a wide range of model sizes. Different modalities exhibit wildly different training dynamics.

Table 5. Uni-Modal scaling law parameters fit to Equation 2 (Chinchilla Scaling Law).

	Code	Image-Text	Image	Molecules	Speech-Text	Speech	Text
A	611.91	320.51	340.96	158.19	180.68	154.45	492.51
B	4484.08	658.31	875.30	189.36	234.13	205.10	1987.40
E	0.16	2.47	2.84	2.39	2.69	3.02	2.42
α	0.37	0.12	0.13	0.37	0.32	0.31	0.18
β	0.32	0.11	0.13	0.26	0.24	0.24	0.22