First Impressions Matter: Predicting Visual Content Popularity through Multimodal Models
Final Project Report for COMS 6720

Dhawal Shah Rohail Alam Shounak Powar

Abstract

In the growing need of social media presence and popularity of becoming an influencer as a fulltime profession, AI can play a major role in assisting in this field. This paper presents a comparative study of three distinct machine learning approaches for predicting social media engagement and content virality, leveraging the large-scale multimodal PixelRec (Cheng et al., 2023) dataset . Addressing the critical challenge of forecasting user engagement with online content, we systematically evaluate: (1) a sophisticated deep multimodal regression model combining SigLIP (Zhai et al., 2023) and DeBERTa (He et al., 2021) encoders with cross-modal attention to predict continuous log-virality scores; (2) a feature-based classification approach utilizing a CatBoost (Prokhorenkova et al., 2019) classifier on a rich set of engineered features for discrete engagement level prediction; and (3) an end-to-end multimodal classification framework integrating BERT (Devlin et al., 2019) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) models to directly classify content into engagement tiers.

1. Introduction

Social media platforms generate massive volumes of user-generated content every day, and understanding what drives user engagement has become increasingly important for content creators, marketers, and platform designers. Engagement metrics such as likes, comments, and shares provide signals about how users interact with content, but accurately predicting these interactions remains challenging due to the multimodal nature of social media posts.

Most existing approaches rely on unimodal features, typically focusing on text or historical interaction data. However, social media posts naturally combine visual content with textual context, and ignoring either modality can lead to incomplete representations. Recent advances in vision–language models enable joint modeling of images and text, offering new opportunities for multimodal engagement prediction.

In this work, we study multimodal engagement prediction using the PixelRec (Cheng et al., 2023) dataset, which provides images, textual metadata, and raw engagement signals including likes, comments, shares, and views. We formulate engagement prediction from three complementary perspectives. First, we model engagement as a five-class classification problem using a fine-tuned vision–language model in a CLIP-style architecture, where image and text encoders are jointly trained and fused through different multimodal interaction strategies. Second, we treat engagement as a continuous regression task and propose a transformer-based multimodal model that predicts a log-transformed engagement score using cross-modal attention and gated fusion. Third, we explore a feature-based classification approach that combines pretrained CLIP (Radford et al., 2021) embeddings with engineered engagement and metadata features and trains a CatBoost (Prokhorenkova et al., 2019) ensemble classifier on the resulting tabular representation.

By evaluating these three approaches within a unified experimental framework, our study provides a systematic comparison of end-to-end deep learning, regression-based modeling, and feature-based ensemble learning for multimodal engagement prediction. This analysis highlights the strengths and limitations of each paradigm and offers practical insights into how different modeling choices affect performance on real-world social media data.

2. Related Work & Background

Predicting user engagement and content popularity on social media platforms has received significant attention due to its importance for content recommendation, digital marketing, and platform design. Engagement is commonly measured using interaction signals such as views, likes, comments, and shares, which act as proxies for user interest and content impact. However, these signals are inherently noisy and influenced by external factors such as platform algorithms, posting time, and audience behavior, making engagement prediction a challenging learning problem.

Early work in this area primarily modeled popularity as a continuous (Khosla et al., 2014) studied image popularity prediction on Flickr by defining popularity in terms of normalized view counts and framing the task as a regression or ranking problem. Their approach combined handcrafted visual features, including color, texture, gradients, object presence, and deep visual descriptors, with social context cues related to the uploader. This work established that image content contributes meaningfully to popularity prediction, but did not consider discrete engagement categories or multimodal textual inputs.

Subsequent research extended popularity prediction to richer multimodal settings. Abousaleh et al. (Abousaleh et al., 2021) proposed a multimodal deep learning framework that integrates engineered visual, social, and temporal features using late fusion to predict continuous popularity scores. While this approach demonstrates the benefit of combining multiple information sources, it relies heavily on handcrafted or platform-specific features and continues to treat popularity as a regression problem over view counts.

Other studies have framed engagement prediction as a classification task. Obučić et al. (Obucic et al., 2023) categorize Facebook post images into low, medium, and high engagement levels based on normalized like counts, using handcrafted visual attributes extracted from external vision APIs and classical machine learning classifiers. This line of work shows that engagement prediction can naturally be modeled as a multi-class classification problem, but typically considers coarse class definitions and image-only inputs.

Recent advances in representation learning have enabled more general multimodal approaches that avoid explicit feature engineering. Vision–language models such as CLIP (Radford et al., 2021) learn aligned representations of images and text using large-scale natural language supervision, while transformer-based architectures like Vision Transformers (Dosovitskiy et al., 2021) and BERT (Devlin et al., 2019) have become standard backbones for visual and textual representation learning. These models allow content-based multimodal representations to be learned directly from data, improving generalization across tasks and domains.

3. Problem Modification

With the original PixelRec (Cheng et al., 2023) dataset as it is, the model would have to be trained as a regression model which is expected to accuractly predict the exact count of the five engagement metrics taken into consideration. This, as one can imagine is a very ambitious task, in fact, too ambitious. To have a well-generalized model that isn't overfit on the training data in such a scenario is near impossible. For this infeasibility of the raw data prediction, a classification approach was considered.

We formulated engagement prediction as a five-class ordinal classification task (*Very Low* to *Very High*). Raw engagement metrics were min-max normalized and summed to create a composite score per post, following the scaling practice common in multimodal learning (Abousaleh et al., 2021). Labels were assigned via quintile-based discretization on the training set to ensure balanced classes without arbitrary thresholds, similar to prior work (Obucic et al., 2023).

Two modeling paradigms were investigated. First, an end-to-end neural approach fused fine-tuned BERT (Devlin et al., 2019) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) encoders using concatenation, attention, or gated fusion. Second, a feature-based method used fixed CLIP embeddings (Radford et al., 2021) augmented with handcrafted features (text statistics, tag encodings, similarity scores) to train a CatBoost classifier. Both were evaluated for capturing ordinal engagement relationships.

Despite further simplifying the problem formulation, the models trained to carry out this classification were still unable to achieve satisfactory performance and were sub-par at best. A key reason for this outcome lies in the absence of a clear and objective definition of what constitutes a "Highly Engaging" post with respect to the engagement indicators considered. Engagement is an inherently multidimensional feature, and there is no universally agreed upon threshold or combination of views, likes, shares, comments, and favorites that reliably distinguishes one engagement class from another. As a result, the class boundaries become ambiguous and noisy, making it difficult for supervised classification models to learn consistent decision rules.

Moreover, discretizing engagement into categorical labels leads to a substantial loss of information. Two posts with nearly identical engagement statistics may be assigned to different classes due to arbitrary thresholding, while posts with vastly different engagement levels may be grouped into the same class. This label ambiguity introduces significant noise into the training process and therefore limits the performance of the model, regardless of the fact that it utilizes state-of-the-art architecture.

Given these limitations, we reverted to a regression-based formulation. However, rather than predicting each engagement signal independently, we further simplified the task by aggregating all engagement indicators into a single continuous `popularity_score`. This approach avoids the need for subjective class definitions while preserving a relative magnitude of engagement across posts.

A given post consists of five engagement values: views, likes, shares, comments, and favorites. Let these be denoted by v, l, s, c and f respectively. The popularity score is computed according to Equation 1

$$\begin{aligned}\texttt{popularity\_score} = &(1 - \mathrm{Cor}(V, L)) \cdot l \\ &+ (1 - \mathrm{Cor}(V, S)) \cdot s \\ &+ (1 - \mathrm{Cor}(V, C)) \cdot c \\ &+ (1 - \mathrm{Cor}(V, F)) \cdot f\end{aligned} \tag{1}$$

Here, $\mathrm{Cor}(\cdot, \cdot)$ denotes the Pearson correlation coefficient computed over the entire dataset. The capitalized variables V, L, S, C and F represent vectors containing all values of views, likes, shares, comments, and favorites across the dataset.

The intuition behind this formulation is that views serve as a baseline and a strong indicator of exposure, while other engagement signals are weighted according to how much additional information they contribute beyond views alone. Engagement types that are highly correlated with views (e.g., likes) have a low weight, as they largely reflect visibility rather than active user interaction. Conversely, signals with lower correlation to views (such as comments or shares) are assigned higher relative importance, as they better capture deeper forms of engagement.

To obtain a more compact and numerically stable target representation, the models are trained to predict the logarithm of the popularity score rather than the raw value. This transformation mitigates the effects of extreme outliers, reduces skewness in the target distribution, and emphasizes relative differences between posts rather than absolute magnitudes. This approach closely mirrors the popularity score formulation used in the SMP dataset (Cheng et al., 2023), which served as the primary motivator for this approach.

4. Methodology

4.1. Fine-Tuned Multimodal Classifier (BERT + ViT)

Data Processing and Labels: We use the PixelRec dataset, where each sample consists of an image, textual metadata (title, tags, description), and raw engagement statistics. Images are resized and normalized using the Vision Transformer image processor. Textual fields are concatenated and tokenized using a pretrained BERT (Devlin et al., 2019) tokenizer with padding and truncation. Engagement supervision is derived using a normalized engagement score:

$$\text{Engagement Score} = \frac{\begin{array}{c}(1 \times \text{likes} + 2 \times \text{comments} + \\ 3 \times \text{shares} + 3 \times \text{favorites})\end{array}}{\max(\text{views}, 1)}.$$

Scores are mapped into five ordered classes (*Very Low* to *Very High*) using global percentile thresholds (20/40/60/80), computed on the training set. Labels are precomputed and used directly during training.

Model and Fusion: The model follows a CLIP (Radford et al., 2021)-style architecture with separate encoders for vision and text. Images are encoded using a pretrained Vision Transformer (Dosovitskiy et al., 2021) (`google/vit-base-patch16-224`), while text is encoded using a pretrained BERT (Devlin et al., 2019) model (`bert-base-uncased`). Both encoders are fine-tuned jointly. Multimodal fusion is performed using one of three strategies: feature concatenation, cross-attention, or gated fusion, allowing the model to capture complementary visual and textual cues. A lightweight classification head predicts one of the five engagement classes.

Training Setup: The model is trained using AdamW with a learning rate of 2×10^{-5}, weight decay 0.01, batch size 16, and one training epoch. Predefined training, validation, and test splits are used. Optimization is performed using multiclass cross-entropy loss.

4.2. Feature-Based CatBoost Classifier (CLIP + Engineered Features

Data Processing: For the CatBoost (Prokhorenkova et al., 2019)-based approach, we use the same PixelRec (Cheng et al., 2023) dataset and engagement labels defined in Section 4.1. Predefined training, validation, and test splits are preserved throughout the pipeline. Engagement labels are converted into integer class indices using label encoding. Images are loaded and converted to RGB format, with a neutral placeholder image used when files are missing. Textual metadata, including the title, tag, and description, is concatenated into a single input string prior to feature extraction.

Feature Extraction: Multimodal representations are extracted using a pretrained CLIP (Radford et al., 2021) model (`openai/clip-vit-base-patch32`). For each post, both image and text inputs are encoded into fixed-dimensional embeddings using the CLIP (Radford et al., 2021) image and text encoders. To explicitly capture cross-modal alignment, the cosine similarity between image and text embeddings is computed and included as an additional feature.

Engineered Features: In addition to CLIP (Radford et al., 2021) embeddings, structured features derived from engagement signals and metadata are incorporated. These include raw engagement counts (likes, comments, shares, views, favorites), their log-transformed variants to reduce skew, and derived ratios such as engagement rate and interaction ratios. Lightweight textual statistics (character length and word count of titles and descriptions) are included, and categorical tags are label-encoded and treated as categorical features by the classifier. All features are concatenated into a single tabular representation.

Model and Training Setup: The final classifier is a CatBoost (Prokhorenkova et al., 2019) multiclass ensemble trained on the combined feature set. CatBoost (Prokhorenkova et al., 2019) is well-suited for heterogeneous, high-dimensional tabular data and natively supports categorical inputs. The model is trained using multiclass cross-entropy loss with a learning rate of 0.03, tree depth of 6, Bernoulli bootstrapping, and ℓ_2 regularization. Training is performed on GPU with early stopping based on validation performance.

4.3. Multimodal Regression-Based Engagement Prediction

Data Processing and Target Construction: For regression, we use the same multimodal inputs from PixelRec. Images are processed using a SigLIP (Zhai et al., 2023) vision encoder, while textual metadata is split into title and contextual text (tags and description) and encoded using pretrained DeBERTa models. The regression target is a continuous log-virality score computed by aggregating views, likes, comments, shares, and favorites using correlation-based weighting, followed by logarithmic transformation to reduce skew and outlier effects.

Model Architecture: The regression model employs a multimodal transformer architecture. Visual and textual features are projected into a shared embedding space and fused using early bidirectional cross-modal attention. Global modality representations are obtained via pooling and combined using a gated fusion mechanism. The fused representation is refined using transformer fusion layers and passed to a regression head to predict a single scalar engagement value.

Training Setup: The model is trained end-to-end using AdamW with a learning rate of 2×10^{-5}, weight decay 0.01, batch size 16, and gradient accumulation over three steps. Training is performed for 50 epochs with mixed-precision (FP16). A cosine learning rate schedule with warmup is used, along with gradient clipping (max norm 1.0). The model is optimized using a hybrid loss combining mean squared error and Huber loss (0.7/0.3).

5. Results

5.1. Fine-Tuned Multimodal Classifier (BERT + ViT)

The dual-encoder model, combining BERT (Devlin et al., 2019) for text and Vision Transformer (ViT) (Dosovitskiy et al., 2021) for images with concatenation fusion, failed to learn a discriminative policy for engagement prediction. As shown in Table 1, the model achieved a test accuracy of 35.94%, which corresponds to the proportion of the majority class ('Very Low') in the test set.

Table 1. Performance of the BERT+ViT multimodal classifier on the test set. The accuracy reflects the prevalence of the 'Very Low' class.

Metric	Value
Test Accuracy	35.94%
Macro Avg. F1-Score	0.11
Weighted Avg. F1-Score	0.19

This result indicates a severe failure in optimization, where the model defaults to predicting the most frequent class to minimize the cross-entropy loss, bypassing any meaningful learning from the multimodal features.

5.2. Feature-Based CatBoost Classifier (CLIP + Engineered Features)

In contrast to the fine-tuned model, the feature-based CatBoost (Prokhorenkova et al., 2019) classifier demonstrated a stable and significantly improved ability to discriminate between engagement classes. This approach leveraged fixed CLIP embeddings augmented with handcrafted metadata and statistical features.

The model's performance, summarized in Table 2, shows consistent results across the validation and test sets, with a test accuracy of **43.18%**. The close agreement between validation accuracy (45.28%) and test accuracy (43.18%) indicates good generalization without overfitting. The robustness of the approach is further confirmed by 5-fold cross-validation, yielding a mean accuracy of 0.4382 ± 0.0029.

Table 2. Performance summary of the feature-based CatBoost classifier. CV denotes 5-fold cross-validation.

Metric	Value
Validation Accuracy	45.28%
Test Accuracy	43.18%
Test F1-Score (Macro)	0.2471
Test F1-Score (Weighted)	0.3855
CV Accuracy (Mean ± Std)	0.4382 ± 0.0029
CV F1-Score, Macro (Mean ± Std)	0.2530 ± 0.0022

The macro F1-score, though modest, confirms that the model makes predictions across all five classes rather than collapsing to the majority. The weighted F1-score, which accounts for class support, is substantially higher. This performance profile suggests that while the classifier successfully learned patterns beyond the trivial baseline, the task remains challenging, with better recall for the more frequent classes.

5.3. Multimodal Regression-Based Engagement Prediction

5.3.1. SigLIP + DeBERTa

The primary point of evaluation for the regression-based approaches is the R^2 metric (coefficient of determination) because it directly measures how well a model explains the variability in the target variable (popularity).

Specifically, R^2 quantifies the proportion of variance in the ground-truth labels that can be explained by the model's predictions. An R^2 value of 1 indicates a perfect fit, where the model explains all observed variability, while a value of 0 indicates that the model performs no better than a baseline predictor that always outputs the mean of the target variable. Negative values further indicate that the model performs worse than this baseline.

Using $R^2$is particularly suitable here because it is scale-independent and allows fair comparison across models that may differ in architecture but are trained on the same target variable. Unlike error-based metrics such as MSE or MAE, which depend on the absolute scale of the target values, R^2 focuses on relative explanatory power, making it easier to assess whether a new model meaningfully improves performance over an existing one. Models are saved only if they achieve a better R^2 value than the previously best-performing model.

With the `google/siglip-so400m-patch14-384` (Zhai et al., 2023) vision encoder and the `microsoft/deberta-v3-large` (He et al., 2021) text encoder, the following training plots are obtained and shown in Figures 1 and 2

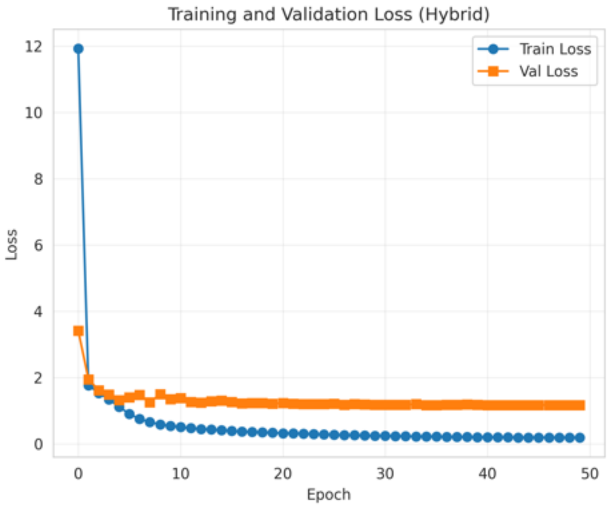

Figure 1. Training and Validation Losses with the hybrid loss function as described in 4.3

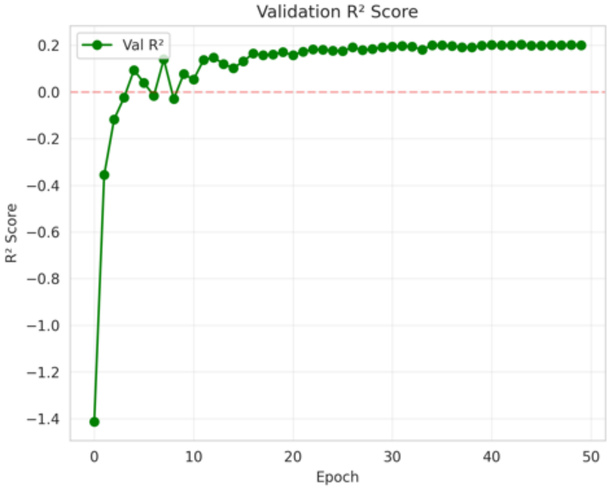

Figure 2. R^2 plot for the same architecture. **Best** $R^2 = 0.2033$

5.3.2. CLIP + DeBERTa

Using the `openai/clip-vit-large-patch14` (Radford et al., 2021) vision encoder along with the same `microsoft/deberta-v3-large` (He et al., 2021) text encoder, the plots depicted in Figure 3 are obtained. The training vs validation curve is not shown for this case due to its similarity with Figure 1, thus reducing redundant and non-conclusive information.

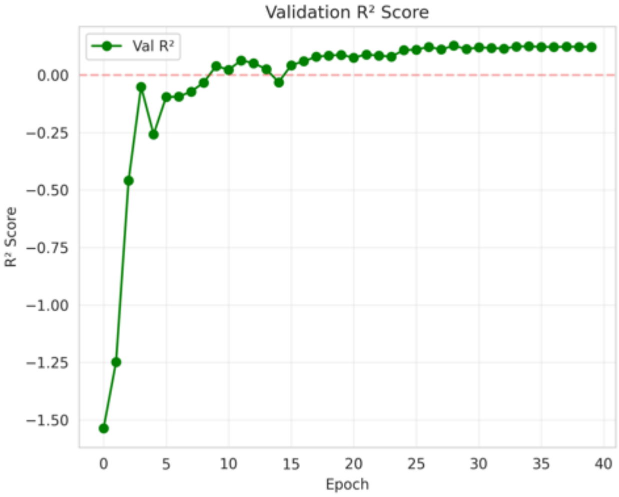

Figure 3. R^2 plot for the same architecture. **Best** $R^2 = 0.1284$

It is evident that the model with the CLIP based vision encoder performs significantly worse than a similar model architecture with the SigLIP vision encoder used instead. A plausible explanation for this difference lies in how the two encoders are trained and the types of representations they tend to learn. CLIP is trained with a contrastive objective that focuses on evaluating images and text relative to each other within a batch. This setup works extremely well for tasks such as retrieval and zero-shot classification, where the goal is to distinguish between categories or concepts. However, it can be less suitable for regression settings, where the model needs to preserve the more fine-grained variations in the input that correspond to the output generation. CLIP embeddings often emphasize high-level semantic distinctions at the expense of this very kind of detailed, continuous information, which can limit the R^2 score.

SigLIP (Zhai et al., 2023), on the other hand, uses a sigmoid-based loss that treats image–text pairs independently rather than comparatively. This tends to produce representations that are better calibrated and retain more sophisticated visual information. For a regression task, this is particularly important, because explaining variance in the target requires features that capture subtle differences between inputs. As a result, the regression head built on top of SigLIP has access to more informative features and is able give a better R^2.

6. Conclusion

This study systematically evaluated three multimodal approaches for predicting social media engagement on the PixelRec dataset. Our findings show that the problem formulation critically impacts success. A fine-tuned BERT-ViT model for five-class classification failed, collapsing to the majority class due to severe imbalance and label noise. A more robust feature-based approach, using CLIP embeddings and engineered features with a CatBoost classifier, achieved a 43.2% test accuracy.

The most promising results came from framing engagement as a continuous regression task. A model fusing SigLIP visual and DeBERTa textual features via cross-modal attention achieved an R² of 0.20. This indicates meaningful explanatory power, though a large portion of engagement variance remains unaccounted for by content alone. A key technical insight was the superiority of the SigLIP encoder over CLIP for regression, as its training objective preserves finer-grained visual information.

In conclusion, engagement on this dataset is more effectively modeled as a continuous target than a classification task. Future work must integrate contextual factors beyond post content and refine fusion techniques to improve predictive performance and practical applicability.

I hope this conclusion effectively wraps up your project report. If you need an abstract or help with formatting the final document, please let me know.

References

Abousaleh, F. S., Cheng, W.-H., Yu, N.-H., and Tsao, Y. Multimodal deep learning framework for image popularity prediction on social media. *IEEE Transactions on Cognitive and Developmental Systems*, 13 (3):679–692, September 2021. ISSN 2379-8939. doi: 10.1109/tcds.2020.3036690. URL http://dx.doi.org/10.1109/TCDS.2020.3036690.

Cheng, Y., Pan, Y., Zhang, J., Ni, Y., Sun, A., and Yuan, F. An image dataset for benchmarking recommender systems with raw pixels, 2023. URL https://arxiv.org/abs/2309.06789.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

He, P., Liu, X., Gao, J., and Chen, W. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

Khosla, A., Das Sarma, A., and Hamid, R. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pp. 867–876, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327442. doi: 10.1145/2566486.2567996. URL https://doi.org/10.1145/2566486.2567996.

Obucic, E., Poturak, M., and Kečo, D. Predicting user engagement of facebook post images in leading universities: A machine learning approach. *Revue d'Intelligence Artificielle*, 37:1139–1145, 08 2023. doi: 10.18280/ria.370426.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. Catboost: unbiased boosting with categorical features, 2019. URL https://arxiv.org/abs/1706.09516.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.