Vision Language Meets Transformers

Anonymous ACL submission

Abstract

Recent vision transformers model images as token sequences but rely on patch-based or continuous pixel embeddings. We extend the vision-language paradigm by representing images as textual sequences derived directly from pixels and modeling them with transformer-based language models. Pixel intensities are mapped to Unicode characters and tokenized using Byte Pair Encoding (BPE), after which a BERT-style transformer is trained on the resulting sequences. As a proof of concept, we pretrain on ImageNet-1K and fine-tune on MNIST and Fashion-MNIST. Results show that vision-language-based transformer models can effectively operate on pixel-derived text and benefit from scalable vocabularies, framing images as a discrete and extensible language.

1 Introduction and Related Work

Recent advances in computer vision have enabled high-accuracy performance across tasks such as image classification, detection, and segmentation, driven by deep learning, large-scale datasets, and increased computational resources (LeCun et al., 2015; Goodfellow, 2016; Bengio et al., 2017). Classical vision pipelines treat images as numeric pixel matrices and rely on handcrafted or learned feature extraction (Ram et al., 2013). The introduction of Convolutional Neural Networks (CNNs) marked a paradigm shift by enabling hierarchical feature learning directly from pixel intensities (Li et al., 2021; O'Shea, 2015; Gu et al., 2018; Yamashita et al., 2018; Zeiler and Fergus, 2014), with residual architectures further improving scalability and performance on complex datasets (Targ et al., 2016).

In parallel, natural language processing (NLP) has advanced rapidly with transformer-based architectures (Vaswani, 2017), which model sequences via self-attention and form the basis of modern language models. Pretrained transformers such as BERT (Devlin, 2018) exhibit strong contextual representations and transferability across tasks. Inspired by these successes, transformers have been adapted to vision by representing images as token sequences. Vision Transformers (ViT) (Dosovitskiy, 2020) and variants such as Convolutional vision Transformers (CvT) (Wu et al., 2021) tokenize images into fixed-size patches, preserving a locality bias from CNNs. More recently, Nguyen et al. (Nguyen et al., 2025) demonstrated that vanilla transformers can operate directly on individual pixels without patch aggregation, challenging the necessity of strong spatial inductive biases. At the intersection of vision and language, multimodal models such as CLIP (Radford et al., 2021) learn joint image–text representations but do not treat image pixels as linguistic units, leaving the question of language-like structure at the pixel level largely unexplored. Moreover, a recent line of work introduced vision-language (Islam et al., 2025), which represents images as textual sequences derived from pixel-level data by mapping pixel intensities to symbolic characters. This enables the use of classical NLP techniques for vision tasks and reveals language-like statistical regularities in pixel-derived sequences. However, the approach remains limited to shallow and lossy architectures, leaving transformer-based language modeling unexplored.

Our Contribution. We extend the vision-language paradigm to transformer architectures by encoding each pixel of grayscale images as a discrete Unicode symbol, yielding a lossless representation with a base vocabulary of 256 tokens. Images are treated as character sequences, and Byte Pair Encoding (BPE) (Shibata et al., 1999) is applied to learn higher-level visual tokens. A BERT-style transformer is trained directly on these sequences to capture contextual dependencies among pixel-derived tokens.

Unlike patch-based or pixel-token vision transformers, our approach adopts a linguistic abstraction: images are modeled as sequences of discrete

symbols rather than continuous embeddings. While ViT-style models are constrained by fixed hidden sizes (e.g., 768) and sequence length, our formulation allows representational capacity to scale with vocabulary size via subword tokenization (e.g., 8K–16K tokens).

As a proof of concept, we pretrain on ImageNet-1K and fine-tune on grayscale MNIST and Fashion-MNIST, demonstrating that vision-language-based transformer models can effectively operate on pixel-derived text and offering a complementary perspective that frames images as a discrete, extensible language.

2 Methodology

In this section, we extend the idea of applying vision-language to transformers. Figure 1 illustrates the whole framework.

2.1 Characterization

Image Normalization and Grayscale Projection. As preprocessing, all images are resized to a fixed resolution of 224×224 pixels and converted to grayscale for color inputs. Let $I \in [0, 255]^{224\times224}$ denote the resulting image. This transformation collapses the RGB channels into a single luminance value, producing a dense matrix of pixel intensities while enforcing a uniform sequence length across the dataset. The grayscale conversion preserves essential structural information relevant for downstream visual modeling while reducing input dimensionality, yielding a simpler and more tractable corpus for training. As illustrated in Figure 1, the MNIST and Fashion-MNIST datasets are already provided in grayscale and therefore directly processed using 256 intensity levels. For other datasets, images are to be preprocessed into grayscale as a prerequisite for applying the proposed framework.

Image-to-Text Characterization. To transform images into a modality compatible with sequence-based language models, we convert each image into a deterministic sequence of Unicode symbols. Each pixel intensity, $p \in \{0, 1, \ldots, 255\}$ is mapped to a unique character in the Unicode Private Use Area (PUA). We define a mapping: $\{0, \ldots, 255\} \rightarrow \{\text{U+E000}, \ldots, \text{U+E0FF}\}$. This results in a closed vocabulary of exactly 256 atomic symbols, each corresponding to a discrete pixel value. The PUA ensures that these symbols do not collide with natural language tokens and remain fully controllable within the framework. Because the mapping operates at the pixel level, considering each pixel as a character, the representation is lossless for grayscale images and perfectly invertible.

Sequence Linearization. The grayscale image matrix I is flattened in row-major order into a sequence of length $50{,}176$: $\big[(I_{1,1}), (I_{1,2}), \ldots, (I_{224,224})\big]$. Thus, each image yields a fixed-length sequence of 50,176 Unicode characters, where each character corresponds to a single pixel. This representation converts the image into a pure text sequence without requiring any learned tokenizer or patch extraction. In contrast to patch-based encodings, this formulation preserves full pixel resolution and enables direct application of sequence models (e.g., masked language modeling or autoregressive objectives) to the raw visual signal.

2.2 Tokenization

We analyze the subword vocabulary learned by the BPE tokenizer trained over the Unicode-encoded image corpus. Since the input consists of sequences over a fixed 256-symbol alphabet (corresponding to grayscale intensities mapped to Unicode code points), BPE operates by merging frequently co-occurring symbol pairs into composite tokens of increasing length. To characterize the structure of the learned vocabulary, we group all tokens by their character length, where a length-1 token corresponds to a single Unicode pixel symbol and tokens of length $\ell > 1$ represent ℓ-character substrings formed through one or more merge operations.

2.3 Transformer Implementation

As illustrated in Figure 1, our approach employs a standard encoder-only transformer architecture, comprising stacked self-attention and feed-forward layers. The proposed vision-language formulation is model-agnostic and can, in principle, be instantiated with any transformer encoder capable of processing token sequences. In this work, we employ BERT as a proof-of-concept to demonstrate the feasibility of modeling pixel-derived linguistic representations using established NLP architectures.

Pretraining. We pretrain a BERT-style transformer encoder on the ImageNet-1K dataset, treating each image as a sequence of BPE tokens. The model is trained using the masked language modeling (MLM) objective, where a subset of visual tokens is randomly masked, and the model is optimized to predict the original tokens based on

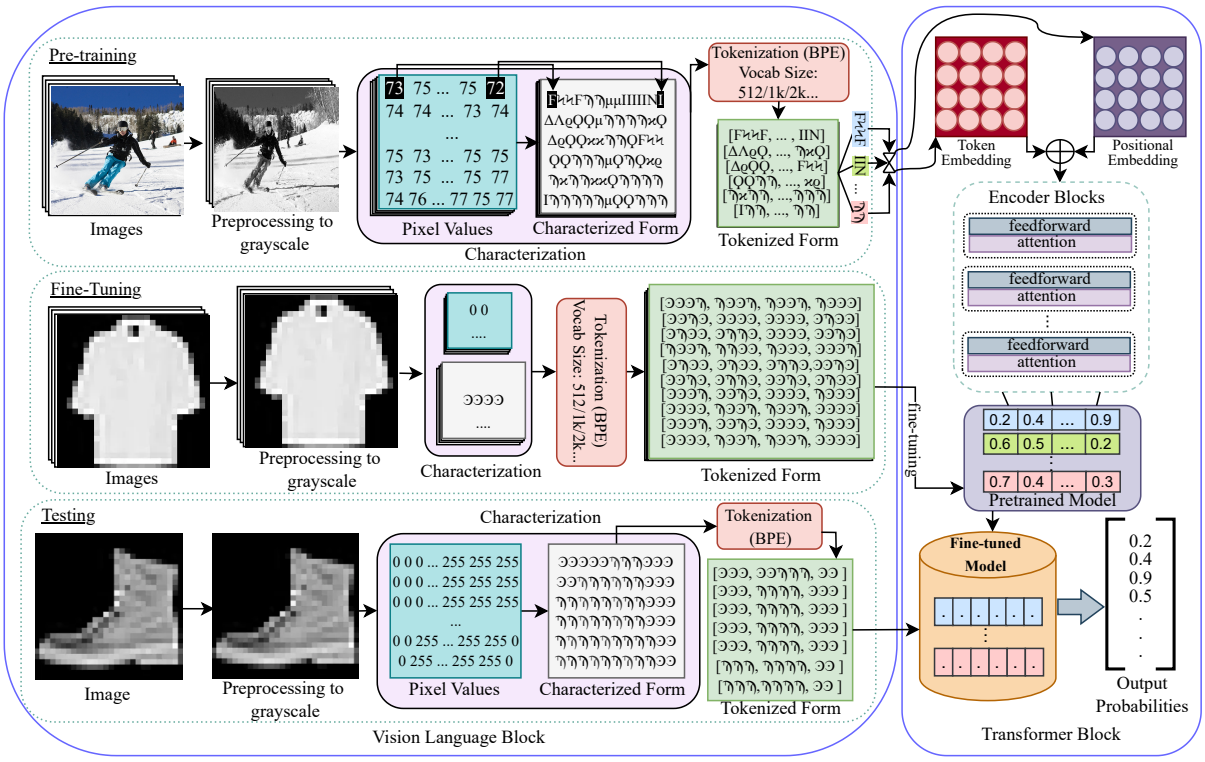

Figure 1: Overview of the vision-language Transformer framework. Images are resized and converted to grayscale, then each pixel intensity is mapped to a unique Unicode character to form a character-level visual text sequence. The resulting sequences are tokenized using Byte Pair Encoding (BPE) and modeled with a transformer trained using a masked language modeling objective, followed by fine-tuning the model.

their surrounding context. Through this objective, the transformer learns contextual dependencies among pixel-derived tokens using self-attention, effectively modeling long-range spatial relationships in the image as linguistic context. No label supervision is used during pretraining.

Fine-tuning and Classification. Following pretraining, the model is fine-tuned on smaller grayscale datasets (MNIST and Fashion-MNIST). Images from these datasets are converted to Unicode sequences using the same pixel-to-character mapping and tokenized using the pretrained BPE tokenizer. We choose image classification as a downstream task. For classification, a task-specific classification head is added on top of the transformer, and the representation of the special `[CLS]` token is used for prediction. Fine-tuning updates all model parameters using labeled data.

This implementation demonstrates that a standard transformer encoder, without architectural modification, can effectively model pixel-level vision-language representations. The use of BERT serves to validate the generality of the framework and establishes a baseline for future exploration with alternative transformer architectures.

3 Experimental Result Analysis

We present a proof-of-concept evaluation of vision-language-based transformer modeling on MNIST and Fashion-MNIST. We compare our lossless approach with the original vision–language lossy baseline (Islam et al., 2025), which reduces image resolution via pixel-intensity quantization (e.g., 8 or 32 levels) and encodes pairs of vertically adjacent pixels as a single character. In contrast, we represent each pixel as a distinct character, preserving full pixel-level granularity and enabling direct application of standard language modeling techniques without spatial aggregation. For fair comparison, all methods use the same BERT-based transformer architecture with hidden dimension 768, ensuring that observed differences stem from representation choices rather than model capacity.

Table 1 reports classification accuracy on MNIST and Fashion-MNIST across different vocabulary sizes and pixel quantization levels, revealing a strong interaction between pixel granularity and vocabulary capacity. For coarse quantization (8 shades), high performance is achieved with small vocabularies (256–512), while larger vocabularies consistently degrade accuracy, indi-

Table 1: Classification accuracy (%) on MNIST (M) and Fashion-MNIST (F-M) across vocabulary sizes and pixel quantization levels. The 8- and 32-shade settings represent two vertically adjacent pixels as a single character, while the 256-shade setting treats each individual pixel as a distinct character.

Vocab. Size	8 Shades (Lossy)		32 Shades (Lossy)		256 Shades (Lossless)	
	M	F-M	M	F-M	M	F-M
256	98.50	87.85	–	–	–	–
512	97.69	86.64	**99.23**	91.77	94.90	83.63
1024	96.70	85.66	99.02	90.89	93.20	83.73
2048	–	–	98.86	90.46	98.00	91.40
4096	–	–	–	–	98.08	**92.32**
8192	–	–	–	–	98.08	91.96

cating that heavily quantized representations do not benefit from increased lexical capacity. With 32 shades, performance improves substantially, especially on Fashion-MNIST, and intermediate vocabularies (512–1024) perform best, suggesting that finer pixel distinctions require moderate vocabulary capacity. In contrast, the 256-shade setting, corresponding to the pixel-level vision-language formulation, exhibits a markedly different trend. Small vocabularies (512–1024) perform poorly, while larger vocabularies are essential for strong performance. Vocabulary sizes of 2048–8192 yield substantial gains, with the best results achieving state-of-the-art–level accuracy of ≈99% on MNIST and ≈93% on Fashion-MNIST. Overall, these results highlight a clear trade-off: coarse representations are effective with small vocabularies, whereas fine-grained, pixel-level vision-language representations require larger lexical capacity to capture richer visual token patterns.

Furthermore, Table 2 summarizes the n-gram token length distributions learned by BPE under the 256-shade pixel representation. The number of unigram tokens remains fixed at 256, corresponding to the pixel-level character inventory. As vocabulary size increases, bigram tokens grow sharply (from 240 at 512 to 6543 at 8192), indicating that frequently co-occurring pixel pairs dominate the learned visual lexicon and that larger vocabularies enable finer modeling of local interactions. Higher-order tokens emerge only with sufficient vocabulary capacity. Trigrams and longer n-grams are rare at small vocabularies but become increasingly prevalent at larger sizes, with the 8192 setting learning tokens spanning up to 256 characters. Although sparse, these long tokens capture recurring higher-order visual structures over extended pixel sequences. Compared to the 4096 setting, the 8192 vocabulary substantially increases both the number and diversity of mid- and long-range tokens. Overall, the distribution exhibits a pronounced long-tail pattern, closely mirroring subword statistics in natural language and explaining the performance gains observed with larger vocabularies in fine-grained pixel-level vision-language modeling.

Table 2: Distribution of n-gram token lengths across different vocabulary sizes for 256 shades.

Token Length	512	1024	2048	4096	8192
Unigram	256	256	256	256	256
Bigram	240	732	1653	3382	6543
Trigram	0	2	10	148	929
4-gram	5	18	104	258	295
5-gram	4	4	4	6	17
6-gram	1	1	1	4	8
7-gram	0	0	0	0	4
8-gram	3	5	8	26	103
12-gram	0	0	1	1	2
16-gram	2	3	5	6	15
20-gram	0	0	0	0	1
24-gram	0	0	0	0	2
28-gram	0	0	0	0	1
32-gram	1	2	3	5	6
40-gram	0	0	0	0	1
44-gram	0	0	0	0	1
48-gram	0	0	0	1	1
64-gram	0	1	2	2	4
128-gram	0	0	1	1	2
256-gram	0	0	0	0	1

4 Conclusion

We extend the vision-language paradigm to transformer-based models by representing images as textual sequences derived directly from pixel-level data. By mapping pixel intensities to discrete symbols, applying subword tokenization, and training a transformer, we show that standard NLP architectures can effectively model pixel-derived visual representations. Pretraining on ImageNet-1K followed by fine-tuning on MNIST and Fashion-MNIST provides a proof of concept that pixel-level vision data exhibits learnable language-like structure. While this study focuses on grayscale images, extending the framework to RGB images, with three channels per pixel, would yield longer and more expressive visual sequences, constituting a richer language. Beyond color images, the approach naturally extends to higher resolutions, diverse visual domains, and tasks such as segmentation, detection, generative modeling, and self-supervised learning. Future work may also explore alternative transformer architectures, efficiency-oriented designs, and structured tokenization strategies, highlighting linguistic abstraction as a flexible foundation for vision modeling.

Limitations

This work is intended as a proof of concept and has limitations. The current experimental evaluation is limited to grayscale datasets and image classification tasks. While the approach generalizes in principle to RGB images and more complex visual domains, additional studies are necessary to assess its effectiveness on higher-dimensional inputs and tasks such as object detection, segmentation, and image generation.

Moreover, our implementation relies on a standard BERT-style encoder with fixed hyperparameters. We do not explore alternative transformer architectures, tokenization strategies, or optimization techniques that may yield better efficiency or performance. Addressing these limitations is an important direction for future research.

References

Yoshua Bengio, Ian Goodfellow, and Aaron Courville. 2017. *Deep learning*, volume 1. MIT press Cambridge, MA, USA.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ian Goodfellow. 2016. Deep learning.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and 1 others. 2018. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377.

Aminul Islam, Md Mustakin Alam, and Shaker Islam. 2025. The birth of vision language. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 12361–12370, New York, NY, USA. Association for Computing Machinery.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019.

Duy Kien Nguyen, Mido Assran, Unnat Jain, Martin R. Oswald, Cees G. M. Snoek, and Xinlei Chen. 2025. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. In *The Thirteenth International Conference on Learning Representations*.

K O'Shea. 2015. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Idan Ram, Michael Elad, and Israel Cohen. 2013. Image processing using smooth ordering of its patches. *IEEE transactions on image processing*, 22(7):2764–2774.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching.

Sasha Targ, Diogo Almeida, and Kevin Lyman. 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31.

Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.