MiniViTex: A Lightweight Vision-and-Text Transformer for Multi-Label Image Caption Classification

Zongxi Qiu [1] Chengjie Deng [1] Shuyu Bai [1]

Abstract

MiniViTex is a novel, lightweight multimodal transformer architecture designed for efficient and accurate multi-label image caption classification. By employing early token-level fusion of visual and textual information, MiniViTex effectively captures complementary semantic cues, enhancing classification performance significantly over image-only approaches. Comparative analyses demonstrate the model's superior balance of accuracy and computational efficiency using compact encoders such as EfficientNetV2-S and BERT-Mini. Empirical evaluations also reveal benefits of larger batch sizes in improving generalization. MiniViTex represents a practical, robust framework for multimodal learning applicable in real-world AI scenarios.

1. Introduction

Real-world image understanding often benefits from leveraging multiple data modalities, such as images and accompanying text. In many settings (e.g. social media posts or news articles), an image is paired with a descriptive caption that can provide complementary cues about the image's content. However, designing models that effectively fuse visual and textual information while remaining efficient is challenging. Multi label image classification – where each image can have a set of correct labels rather than a single label – is particularly demanding due to the diversity of concepts per image. Incorporating textual captions into such classifiers has the potential to disambiguate visual features and improve predictive accuracy, as the text may highlight salient objects or contexts that are less obvious from the image alone. At the same time, deploying multimodal models in practice requires keeping model size and inference time in check, motivating research into lightweight vision language architectures.

Recent advances in vision-and-language learning have yielded powerful but resource intensive models. For example, the CLIP model by Radford et al. (2021) aligns images and text in a joint embedding space using separate encoders for each modality trained on hundreds of millions of image-caption pairs. While such large-scale dual-encoder models achieve remarkable zero-shot performance, their hefty computational requirements make them less practical for domain-specific tasks or on-device applications. On the other hand, single-stream transformer models like VisualBERT (Li et al., 2019) process visual and textual inputs together in a unified Transformer, enabling rich interactions between modalities. These approaches often build on BERT-like architectures (Devlin et al., 2019) and require heavy visual feature extractors (e.g. a pre-trained ResNet CNN) to encode images, resulting in large model sizes and slow inference. There is a clear need for a unified multimodal model that is compact and efficient without sacrificing too much accuracy.

In this paper, we introduce MiniViTex (Mini Vision-and-Text Transformer), a lightweight multimodal Transformer model designed for multilabel image caption classification. MiniViTex integrates an image encoder and a text encoder into a single unified architecture that jointly processes both modalities from an early stage. The model supports interchangeable visual backbones – including (i) a Convolutional Neural Network (CNN) based approach using EfficientNetV2-S with an attention pooling mechanism, (ii) a classic ResNet-50 CNN with attention pooling, and (iii) a pure Vision Transformer backbone using the DeiT architecture, to exact image tokens, paired with a compact text encoder (BERT-Mini) for caption processing. The extracted visual and textual features are early fused at the token level: MiniViTex adds a learnable '[CLS]' token and feeds the image and text token embeddings together through a Transformer encoder, enabling cross-modal attention interactions. By using a shared positional encoding for the concatenated token sequence, the model treats image patches (or pooled image features) and words equivalently in the Transformer layers. This design is inspired by recent efficient V&L models like ViLT, which prove that a single transformer can handle both modalities with minimal per-modality processing overhead. Unlike large prior models, our approach

[1] SID: 520005325, 520030671, 520204984. Correspondence to: <{zqiu6882, cden5509, sbai4644}@uni.sydney.edu.au>.

emphasizes model compactness – for instance, our text encoder is a 4-layer, 256-dimensional BERT-Mini (Bhargava et al., 2021; Turc et al., 2019), which has orders of magnitude fewer parameters than BERT-base, and our vision backbones are chosen for their favorable accuracy-efficiency trade offs.

We evaluate MiniViTex on the Multi Label Classification Competition 2025 dataset, which consists of images each annotated with one or more labels and accompanied by a brief caption. The task is to predict the set of labels for each image. We report standard multi-label evaluation metrics including mean Average Precision (mAP), Micro F1, Macro F1, and mean sample wise F1. Our results demonstrate that MiniViTex achieves strong multi-label classification performance, outperforming comparable single modal baselines. In particular, we find that incorporating the caption yields a significant improvement in F1 scores over using the image alone, confirming the value of multimodal fusion. An ablation study shows that removing the text input causes a substantial drop in mean F1, underscoring how the language channel provides complementary information. We also compare the three supported image encoders and highlight the trade-offs between them: e.g., the convolutional backbones (ResNet-50, EfficientNetV2-S) vs. the pure Vision Transformer (DeiT-S) differ in feature representation and efficiency, each influencing the overall model's accuracy and speed.

Contributions: In summary, this work makes the following key contributions:

1. **MiniViTex Architecture:** We propose MiniViTex, a unified Transformer-based model that processes images and text jointly for multi-label classification. The architecture is lightweight – leveraging a small pretrained text encoder and efficient vision backbones – making it suitable for real-world applications with limited computational resources.
2. **Multimodal Fusion Effectiveness:** We demonstrate that early token-level fusion of visual and textual features significantly boosts classification performance. MiniViTex yields higher mAP and F1 scores than image-only models on the benchmark dataset, illustrating the complementary strengths of the caption and image modalities.
3. **Flexible and Efficient Backbones:** Our framework supports multiple image encoders (EfficientNetV2-S, ResNet-50, and DeiT-S) under a consistent fusion model. We provide an extensive comparison of these backbones, highlighting how modern efficient CNNs and vision transformers perform in a multimodal context, and we analyze the accuracy–complexity trade-offs to guide model selection for deployment.
4. **Ablation and Analysis:** Through ablation studies, we quantify the impact of each component of MiniViTex. We show, for example, that removing the text input degrades the mean sample-wise F1 by a large margin, confirming the importance of multimodal learning. By combining efficiency and efficacy, MiniViTex aims to bridge the gap between heavyweight vision-language models and practical multi-label classification needs. To the best of our knowledge, this is one of the fewer works that explicitly targets a lightweight unified model for image+caption multilabel classification, and we hope that it encourages further research into compact multimodal architectures for real-world AI systems.

2. Related Work

2.1. Vision-Language Transformers

The past few years have seen rapid progress in models that jointly reason over images and text. Early vision-language transformers often employed a two-stream (dual encoder) design or relied on pre-extracted region features. For instance, ViLBERT (Lu et al., 2019) and LXMERT (Tan & Bansal, 2019) use separate CNN and language encoder streams with cross-attention to connect visual regions and words. In contrast, single-stream models like VisualBERT and UNITER concatenate image and text tokens from the start and feed them into a unified Transformer encoder (Li et al., 2019). VisualBERT in particular is a seminal single-stream model that integrates BERT with visual region features: it takes object detections from a CNN (e.g. Faster R-CNN) as "visual tokens" and processes them alongside word tokens through BERT's layers (Devlin et al., 2019). This approach allows the model to learn intricate alignments between language and specific image regions through self-attention, and it achieved strong results on tasks such as VQA and caption-based retrieval while being conceptually simpler than dual-stream architectures (Chen et al., 2020). However, VisualBERT and similar models still depended on powerful CNN backbones (with hundreds of millions of parameters) to generate region features, meaning the overall system was far from lightweight.

Recent research has increasingly emphasised efficiency in vision-language models. Kim et al. (2021) introduced ViLT, a Vision-and-Language Transformer that drastically simplifies the visual processing by removing heavy CNNs and region proposals. ViLT feeds raw image patches (projected linearly) and text into a shared Transformer, essentially letting the Transformer learn visual features from scratch alongside language features. This "CNN-free" design yields a much smaller and faster model, demonstrating that transformers can directly handle pixel inputs when given sufficient training. Although ViLT's image understanding lags behind models using pre-trained CNN features, it proved that significant speedups ($4 - 10\times$ faster inference) are possible in multimodal models by using a single, unified Transformer for both. Our MiniViTex shares ViLT's philosophy of early fusion in one transformer, but we take a hybrid approach: we leverage lightweight pre-trained vision backbones (rather than learning from raw pixels) to retain strong image recognition capabilities, while still benefiting from a unified Transformer fusion that keeps params low.

Another line of work focuses on contrastive dual-encoder models. The most prominent example is OpenAI's CLIP (Contrastive Language–Image Pretraining) model (Radford et al., 2021). CLIP consists of an image encoder and a text encoder that are trained jointly on large-scale web data to produce aligned representations – images and their paired captions are mapped to nearby vectors in a shared embedding space, while mismatched image-caption pairs are mapped far. Notably, CLIP replaces direct image caption prediction with a simpler objective of predicting the correct pairings in a batch via a contrastive, which proved highly effective for learning transferable features. After training on 400 million image-text examples, CLIP's encoders can be used for zero-shot classification by comparing image features to text features of class names (performing near ImageNet accuracy without fine-tuning). While powerful, CLIP's approach uses two large model streams (a Vision Transformer or CNN and a Transformer language model) and thus has a high computational footprint, making it challenging to fine-tune or deploy on smaller tasks. In our work, we aim for a single-stream model that can be trained on a modest-sized dataset from scratch (or with minimal pretraining) and still benefit from image-text synergy. Unlike CLIP's late fusion (similarity at the embedding level), we perform early fusion – which has been shown to be advantageous when one needs to predict task-specific labels rather than generic embeddings (Li et al., 2019). Early fusion allows our classifier to directly attend across image patches and words, potentially learning more task-relevant interactions (e.g. a particular word in the caption focusing attention on the corresponding part of the image that relates to a certain label).

2.2. Backbones for Vision and Text Inputs

Our model builds on advances in both visual and textual representations. For vision, Convolutional Neural Networks (CNNs) have long been the backbone of image recognition. ResNet-50 (He et al., 2016) is a classic CNN with skip-connection architecture that achieves strong accuracy on ImageNet with 25 million parameters, and it has been a popular choice for extracting visual features in multimodal (Ilharco et al., 2021). More recently, EfficientNet and its successor EfficientNetV2 (Tan & Le, 2021) have set new standards for parameter efficiency in CNNs. EfficientNetV2 employs neural architecture search and scaling strategies to achieve higher accuracy with fewer FLOPs, resulting in models that train faster and are up to 6.8 times smaller than prior CNNs for similar performance. We include EfficientNetV2-S as one option for MiniViTex's image encoder, expecting it to provide strong image features at relatively low cost. Meanwhile, the paradigm is shifting toward Vision Transformers (ViT), inspired by the success of Transformers in NLP. ViT models treat an image as a sequence of patch embeddings and have shown that with enough data, they can surpass CNNs in image classification accuracy (Dosovitskiy et al., 2021). DeiT (Data-Efficient Image Transformers by Touvron et al. (2021)) made ViTs more accessible by demonstrating that one can train a ViT on ImageNet-1k (without external data) using distillation and optimisation tricks. The DeiT-S variant (Small) is a 22M-parameter transformer that attains competitive ImageNet results in under 3 days of training (Hugging Face, n.d.). We leverage DeiT-S as a representative transformer-based vision backbone in our experiments, to compare against CNN-based backbones. By evaluating MiniViTex with ResNet-50, EfficientNetV2-S, and DeiT-S in the same framework, our work sheds light on how different visual feature paradigms perform in a multimodal classification setting. Prior studies of (Strudel et al., 2021) have also explored transformer backbones for multilabel classification; our contribution is to examine them in the context of joint image-text modelling.

Attention mechanisms have also been integrated into CNN-based frameworks to overcome limitations associated with global pooling methods. In traditional CNN classifiers, after a series of convolutional layers producing a spatial feature map, a global pooling operation (e.g., global average pooling) is often used to aggregate features before the final prediction layer. While simple and effective, global average pooling treats all spatial locations equally, which may be suboptimal for complex scenes or multi-label images where multiple objects/features are present. Attention-based pooling mechanisms have been proposed to allow the model to learn which parts of the image are most important. For instance, Ilharco et al. (2021) introduced an attention pooling approach in a multiple-instance learning scenario, showing that a trainable attention mechanism can weight instance-level features and improve prediction by focusing on critical instances. In the context of CNNs, an attention pooling layer can take the set of feature vectors (one per spatial location or region) and produce a weighted sum, where the weights are computed by an attention network that emphasizes relevant regions. This yields a more informative global descriptor than unweighted averaging, especially for multi-label classification where the image might contain several relevant regions that each correspond to different labels.

For text, the dominant approach is based on the Transformer encoder from BERT (Devlin et al., 2019). BERT-base (12 layers, 768d) revolutionized NLP by providing a powerful pre-trained language representation, but its size can be prohibitive for lightweight applications. Subsequent research introduced compact BERT variants through knowledge distillation and architectural scaling. Bhargava et al. (2021) released a family of pre-trained BERT Mini models with significantly fewer parameters, including a 4-layer, 256-dimensional model termed BERT-Mini. Despite having only 11 million parameters, BERT-Mini retains a surprising amount of performance on language understanding tasks (e.g., 66 GLUE score vs 77 for BERT-base) while being much faster to run. Such compact models are ideal for multimodal tasks where the language input (captions) is relatively short and domain-specific, and where using a full-sized BERT would dominate the parameter count. In our design, we choose BERT-Mini as the text encoder to keep the model lightweight. We initialize it with pre-trained weights so that even with its small size, it provides a good language understanding prior for the model. Similar strategies of using smaller transformers have been employed in mobile NLP and multimodal systems (e.g., TinyBERT and DistilBERT), but our work integrates a mini transformer on the text side with an efficient image encoder on the vision side in a unified architecture.

2.3. Multimodal Fusion Approaches

Combining visual and textual representations is at the heart of vision-language research. A variety of architectures have been explored to fuse information from the two modalities. Earlier approaches often used two-stream models, where the image and text are processed by separate sub-networks that then interact via a fusion layer. For example, ViLBERT (Lu et al., 2019) and LXMERT (Tan & Bansal, 2019) employ two Transformer encoders – one for vision, one for language – and perform cross-modal attention at a later stage to allow the modalities to exchange information. These models were typically pre-trained on large image-caption datasets with proxy tasks (like matching or masking) and have been successful on tasks like VQA and captioning. However, two-stream models can be heavy, since they effectively double the number of parameters (one transformer per modality) and require additional layers for cross-attention.

In contrast, single-stream (early fusion) models aim to embed both modalities into a single shared transformer. VisualBERT (Li et al., 2019) is a seminal example: it takes region-based visual features (e.g., from a pre-trained Faster R-CNN) and text tokens, adds a special separator token, and feeds the concatenated sequence into a BERT-like Transformer. The model learns alignments between regions and words implicitly through self-attention and was shown to be effective on various vision-language tasks. Similarly, Unicoder-VL and UNITER (Chen et al., 2020) followed this paradigm with different pre-training tasks, all demonstrating the benefit of joint encoding. More recently, the trend has moved toward removing reliance on external region proposals and enabling end-to-end training. ViLT (Kim et al., 2021) is a model that cuts out the convolutional region detector entirely: it patches the image (like ViT) and feeds patch embeddings plus text token embeddings into a unified Transformer. By not using a CNN for the image (aside from a linear projection of patches), ViLT significantly reduces the computational overhead, and it showed that a pure Transformer can learn vision-language tasks given sufficient training. However, ViLT still inherits the relatively high computation cost of attending over many patch tokens, and it needs careful optimisation to perform well.

Another interesting line of work is improving how the fusion happens within a Transformer. The TokenFusion method proposes dynamically identifying uninformative tokens in one modality and replacing them with tokens from the other modality during transformer encoding. The idea is to mitigate the dilution of attention across too many tokens by actively mixing modalities token-by-token. While our approach does not implement TokenFusion's dynamic token replacement, it shares the philosophy of tightly integrating the modalities rather than keeping them separate.

2.4. Multi-Label Image-Text Classification

Multi-label image-text classification remains a relatively under-explored but practically significant task in the field of multimodal learning, where models need to reason jointly about visual content and relevant textual descriptions in order to predict multiple semantic labels. Early approaches have relied heavily on image-only architectures such as ResNet with sigmoid classifiers or graph-based models (e.g., ML-GCN), treating the task as a traditional visual multi-label classification problem. However, due to limited visual cues, these models often suffer from semantic ambiguity in visually similar categories and tend to perform poorly in long-tailed scenarios.

In contrast, text-only approaches use pre-trained language models such as BERT to perform classification based on captions or metadata. This approach is lightweight and effective when the captions are informative, but its performance degrades significantly when the text input is ambiguous, incomplete, or fails to fully describe the image. In order to fully utilize both modalities, late fusion strategies have been proposed, where image and textual representations are independently encoded and concatenated prior to classification. Despite their simplicity, these methods lack fine-grained cross-modal interactions and often struggle to resolve referential ambiguity.

Recent advances employ early fusion architectures inspired by single-stream visual language transformers such as VisualBERT, UNITER, and ViLT.These models integrate image and text tokens at the input level, enabling token alignment and joint representation learning. However, these models typically require large-scale pre-trained backbones that are computationally expensive, limiting their applicability in resource-limited environments. In addition, few of these models explicitly focus on classification as a training goal, especially label co-occurrence modeling and label sparsity robustness.

In short, existing approaches either ignore the synergies between modalities or incur high computational overheads, and few are specialized for lightweight, label-rich, or long-tailed multilabel scenarios. These limitations motivate us to design compact and efficient architectures, such as our proposed MiniViTex, that aim to bridge this gap through efficient early fusion and semantic alignment.

3. Methodology

3.1. Model Overview

MiniViTex is a unified multimodal Transformer architecture that takes an image and its caption as input and outputs a set of predicted labels for the image. The design consists of three main components: (1) an Image Encoder that extracts visual features from the input image, (2) a Text Encoder that produces embeddings for the caption text, and (3) a Multimodal Transformer Fusion module that joins the two modalities and produces the final classification output. Unlike approaches that fuse modalities only at the final prediction stage, MiniViTex performs early fusion at the token level – the image and text representations interact within Transformer layers, allowing the model to attend to relevant image regions and caption words jointly. This section details each component and the training setup.

3.2. Image Encoder

MiniViTex supports flexible image encoders, allowing either convolutional backbones with attention-based tokenization or transformer-based architectures such as DeiT. Regardless of the specific encoder, all visual features are transformed into a unified token sequence of the form:

$$\mathbf{I} = [\mathbf{x}_{\text{cls}}, \mathbf{x}_1, \dots, \mathbf{x}_N] \in \mathbb{R}^{(N+1)\times D}$$

where $\mathbf{x}_{\text{cls}}$ is a global pseudo-CLS token, $\mathbf{x}_i$ are patch-level tokens, and $D = 384$ is the shared multimodal embedding dimension. To ensure stability and prevent token-level feature specialization, we apply a LayerNorm before projecting visual features into the fusion space.

3.2.1. CNN + Attention Pooling:

Given an input image $\mathbf{X} \in \mathbb{R}^{3\times384\times384}$, we use a pretrained convolutional backbone f_{cnn} in `EfficientNetV2-S` or `ResNet-50` to extract spatial feature maps:

$$\mathbf{F} = f_{\text{cnn}}(\mathbf{X}) \in \mathbb{R}^{C\times H\times W},$$

where $C = 1280$ (EfficientNetV2-S) or 2048 (ResNet-50), and $H = W = 12$. The feature map is flattened into $N = H \times W = 144$ patch embeddings:

$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_N] \in \mathbb{R}^{N\times C}.$$

A global summary token is computed by averaging across all spatial positions:

$$\mathbf{z}_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i, \quad \mathbf{Z}' = [\mathbf{z}_{\text{mean}}; \mathbf{z}_1; \dots; \mathbf{z}_N].$$

We then add learnable positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1)\times C}$ and perform one round of multi-head (4 heads with 2 layers) self-attention to enable token interaction:

$$\mathbf{H}_0 = \mathbf{Z}' + \mathbf{E}_{\text{pos}}, \quad \mathbf{H}_1 = \text{MHA}(\mathbf{H}_0).$$

To prevent sharp activation distributions and stabilise training, we apply Layer Normalisation and finally project the normalised features into the multimodal embedding space:

$$\mathbf{H}_2 = \text{LN}(\mathbf{H}_1), \quad \mathbf{I} = \mathbf{H}_2 \cdot \mathbf{W}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{C\times D}.$$

By using attention pooling with CNN features, we aim to give the model a similar capability to concentrate on important visual signals. Such techniques are related to the broader area of visual attention and have been found beneficial in tasks like fine-grained recognition and image captioning where not all parts of an image are equally relevant (Dosovitskiy et al., 2021). It finally produce ViT-style token representations suitable for transformer-based fusion.

3.2.2. Vision Transformer

As a transformer-native encoder, DeiT-S accepts an image $\mathbf{X} \in \mathbb{R}^{3\times224\times224}$ and produces a token sequence via non-overlapping patch embedding and self-attention. The model outputs:

$$\mathbf{P} = [\mathbf{p}_{\text{cls}}, \mathbf{p}_1, \dots, \mathbf{p}_N] \in \mathbb{R}^{(N+1)\times d}, \quad \text{with } d = 384.$$

To ensure consistent feature scaling before fusion, we apply a post-hoc LayerNorm and optionally a linear projection if the hidden size differs from 384:

$$\tilde{\mathbf{P}} = \text{LN}(\mathbf{P}), \quad \mathbf{I} = \tilde{\mathbf{P}} \cdot \mathbf{W}_{\text{proj}}, \quad \mathbf{W}_{\text{proj}} \in \mathbb{R}^{d\times D}.$$

This normalization step mitigates internal covariate shift and ensures token stability across different input domains or image augmentations. The resulting token sequence $\mathbf{I}$ is structurally aligned with that of the CNN encoders and can be directly passed to the multimodal transformer.

3.3. Text Encoder

To represent the caption modality, MiniViTex adopts a compact Transformer-based language encoder built on top of the 4-layer pretrained BERT-Mini model (Turc et al., 2019). This lightweight design is particularly suitable for the multi-label caption classification setting, where the input text is short (typically fewer than 20 words per caption) and the vocabulary domain is relatively narrow. Compared to larger BERT variants, BERT-Mini achieves a significantly lower memory footprint while preserving sufficient language understanding for downstream multimodal tasks. Its simplicity also facilitates faster training and inference, aligning well with MiniViTex's efficiency-oriented design.

Formally, let the caption for an image be tokenized into a sequence of T subword tokens $\{w_1, w_2, \dots, w_T\}$. These are converted into token IDs $\mathbf{t} = [t_1, t_2, \dots, t_T] \in \mathbb{N}^T$, and passed to the BERT encoder along with an attention mask $\mathbf{m} \in \{0, 1\}^T$ indicating valid tokens:

$$\mathbf{H}_{\text{text}} = \text{BERT-Mini}(\mathbf{t}, \mathbf{m}) \in \mathbb{R}^{T \times d_{\text{text}}},$$

where $d_{\text{text}} = 256$ is the hidden size of the pretrained BERT-Mini model. The output $\mathbf{H}_{\text{text}}$ is a sequence of contextualized token embeddings. To align with the multimodal fusion transformer's hidden size $D = 384$, we apply a normalisation and linear projection to each token embedding:

$$\mathbf{H}_{\text{proj}} = \text{LN}(\mathbf{H}_{\text{text}}), \quad \mathbf{E}_{\text{text}} = \mathbf{H}_{\text{proj}} \cdot \mathbf{W}_{\text{proj}} \in \mathbb{R}^{T \times D},$$

where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{d_{\text{text}} \times D}$ is a learnable projection matrix. The result $\mathbf{E}_{\text{text}}$ is the sequence of caption token embeddings used for multimodal fusion.

Unlike many traditional text encoders that use a CLS token to represent the entire sentence, MiniViTex performs fusion with a separate multimodal CLS token. Thus, the original text CLS embedding is treated as a regular token and included in the full sequence:

$$\mathbf{E}_{\text{text}} = [\mathbf{e}_1, \dots, \mathbf{e}_T],$$

with no special classification role. This approach allows all text tokens to participate equally in cross-modal attention and avoids redundancy.

The use of BERT-Mini ensures that the text encoder remains compact (approximately 11M parameters), while LayerNorm and projection stabilize the feature scale and enable smooth integration with image tokens. Given the brevity and structured nature of captions in our dataset, we find that BERT-Mini strikes an optimal balance between representation power and computational efficiency. In practice, it yields comparable performance to heavier encoders such as DistilBERT or BERT-base on our task, with lower memory and latency.

3.4. Cross-Modal Transformer

After obtaining the tokenized visual features $\mathbf{I} \in \mathbb{R}^{(N+1) \times D}$ and text features $\mathbf{E}_{\text{text}} \in \mathbb{R}^{T \times D}$, MiniViTex concatenates them into a single multimodal sequence along with a learnable classification token. This sequence is processed by a stack of Transformer encoder layers that jointly model intra-modal and inter-modal interactions through self-attention.

3.4.1. Token Concatenation and Positional Embedding

We define a learnable multimodal classification token $\mathbf{z}_{\text{cls}} \in \mathbb{R}^D$, shared across all samples. For a given batch of size B, the final input sequence is:

$$\mathbf{X}_0 = [\mathbf{z}_{\text{cls}}; \mathbf{I}; \mathbf{E}_{\text{text}}] \in \mathbb{R}^{B \times (1+N+T) \times D}.$$

To encode token positions, we learn a set of positional embeddings $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{L_{\max} \times D}$, where $L_{\max} \geq 1 + N + T$. We apply position encodings uniformly across image and text tokens:

$$\mathbf{X}_0 = \mathbf{X}_0 + \mathbf{E}_{\text{pos}}[:, : 1 + N + T, :].$$

This unified positional space encourages early alignment between modalities and allows the model to treat all tokens in a coherent embedding geometry. Such alignment has been shown to enhance fine-grained correspondence learning in vision-language models (Li et al., 2019; Kim et al., 2021).

3.4.2. Multimodal Transformer Fusion

The fused sequence $\mathbf{X}_0$ is processed by a stack of L Transformer encoder layers. Each layer consists of multi-head self-attention and feedforward submodules with residual connections and layer normalization:

$$\mathbf{X}_\ell = \text{TransformerLayer}(\mathbf{X}_{\ell-1}), \quad \ell = 1, \dots, L.$$

We again use 4 heads with 2 layers with a feedforwards dimension of 1024. Each attention head in the self-attention mechanism has full access to the entire token sequence, including the CLS, all image patch tokens, and all text tokens. This allows the model to, for example, relate a specific noun phrase to the corresponding visual region, or associate a caption's temporal modifier (e.g., "at night") with relevant visual semantics (e.g., darkness in sky patches).

The early fusion strategy—feeding all tokens together into a unified Transformer from the beginning—enables direct and dynamic cross-modal interaction at every layer. Compared to late fusion (e.g., averaging final embeddings from each modality), early fusion allows richer multi-hop reasoning (Chen et al., 2020; Kim et al., 2021) and better alignment at the token level:

- Visual tokens can attend to semantically aligned words in context (e.g., image patch attending to "cat");

- Text tokens can adjust their representation based on visual evidence (e.g., “park” interpreted as greenery vs surname);
- The global [CLS] token aggregates both modalities to represent the entire (image, caption) pair for prediction.

Formally, the final output of the transformer is:

$$\mathbf{X}_L = [\mathbf{z}_{\text{cls}}^{(L)}; \ldots], \quad \mathbf{z}_{\text{cls}}^{(L)} \in \mathbb{R}^{B \times D}.$$

By retaining all visual patch tokens (not collapsing them via average or attention pooling) and all textual tokens (not compressing into a single vector), MiniViTex allows detailed multimodal interactions to emerge naturally during attention. This token-level design provides several advantages:

- Enables fine-grained alignment, where different parts of the caption (e.g., objects, modifiers, relations) interact with different image regions;
- Facilitates interpretability, as attention weights can be visualized between image patches and specific words;
- Improves robustness in long-tail categories, where weak visual cues may be clarified by textual grounding;
- Leverages the inductive bias of transformers for multi-token fusion, which has been shown effective in ViLT and VisualBERT.

Compared to models that fuse at the embedding or logits level, our approach supports richer reasoning and reduces modality mismatch. This structure also enables ablation at token level and future extension to spatial grounding or generative tasks.

3.5. Classification Head

To predict the presence of each label, MiniViTex applies a lightweight classification head to the final multimodal CLS token representation produced by the transformer. This head consists of a layer normalisation followed by a linear projection:

$$\mathbf{z}_{\text{logits}} = \text{LN}(\mathbf{z}_{\text{cls}}^{(L)}) \cdot \mathbf{W}_{\text{cls}} + \mathbf{b}_{\text{cls}} \in \mathbb{R}^{C},$$

where $\mathbf{z}_{\text{cls}}^{(L)} \in \mathbb{R}^{D}$ is the final hidden state of the multimodal CLS token, $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{D \times C}$, and C is the number of classes.

The output $\mathbf{z}_{\text{logits}}$ consists of raw unnormalized scores (logits), which are not passed through a sigmoid activation within the model forward pass. Instead, we apply a binary cross-entropy loss with logits during training (e.g., PyTorch’s BCEWithLogitsLoss), which internally applies the sigmoid operation:

$$\mathcal{L} = -\frac{1}{C} \sum_{j=1}^{C} \left[y_j \cdot \log \sigma(z_j) + (1 - y_j) \cdot \log(1 - \sigma(z_j)) \right],$$

$$\sigma(z_j) = \frac{1}{1 + e^{-z_j}}.$$

At inference time, we convert logits to probabilities using the sigmoid function:

$$\hat{y}_j = \sigma(z_{\text{logits},j}) \in (0, 1),$$

and apply a fixed threshold $\tau \in (0, 1)$ to determine binary predictions:

$$\tilde{y}_j = \begin{cases} 1 & \text{if } \hat{y}_j > \tau \\ 0 & \text{otherwise} \end{cases}.$$

This threshold can be class-agnostic or class-specific, tuned based on validation F1 or mAP. In our experiments, we use a default threshold of 0.5 unless otherwise noted. This design allows the model to flexibly score each label independently and to handle multiple simultaneous positive classes per instance, which is essential in multi-label classification tasks.

3.6. Model Summary and Design Considerations

In all MiniViTex experiments, we adopt a consistent and lightweight configuration to ensure comparability across backbones and to encourage architectural stability. Specifically:

- The input image resolution is fixed to $384 \times 384 \times 3$, yielding a 12×12 patch grid for CNN-based backbones after downsampling;
- Captions are tokenized and truncated to a maximum sequence length of 32 tokens;
- All token representations, visual, textual, and fused are embedded into a shared hidden dimension $D = 384$, which strikes a balance between expressive capacity and efficiency.

To improve training stability and feature consistency across modules (CNN → Attention → projection, BERT → projection, and Transformer CLS → Logit), we apply layer normalisation extensively throughout the model. This includes post-encoder LayerNorms before projection, within transformer blocks, and at classification heads. We find that such normalisation reduces activation divergence between modalities and simplifies convergence, especially under mixed-precision and large-batch settings.

While we apply a dropout rate of 0.1 within the attention and feedforward submodules of all Transformer layers, we deliberately omit dropout between major components (e.g., between encoder and fusion layers). Empirically, we found this inter-block dropout introduces instability in early fusion settings, it might due to its unstructured noise on token embeddings that are shared across modalities.

Instead, we rely on strong weight decay as the primary form of structural regularisation. Specifically, we use AdamW with a weight decay of 0.1 applied uniformly to all attention, transformer, and classification parameters. Recent studies (Kobayashi et al., 2024; Liu et al., 2023) have shown that high weight decay can act as a powerful regulariser in transformer-based models, promoting generalisation and reducing reliance on dropout noise, particularly when training data is not massive.

The combination of moderate embedding size, deep normalisation, conservative dropout, and aggressive weight decay makes MiniViTex both efficient and robust, enabling us to train stable multimodal models across varying backbones without extensive hyperparameter tuning.

3.7. Dataset and Preprocessing

We evaluate MiniViTex on the Multi-label Classification Competition 2025 dataset, released on Kaggle as part of an in-course challenge (Lin, 2025). The dataset consists of 30,000 training samples, each comprising an image and a short caption, with an average caption length of 13 tokens. There are 18 target labels (1 to 19 except 12) across all samples, with class distributions exhibiting high imbalance: The most dominant class (label 1) appears in approximately 49% of all training samples, with over 22,000 occurrences. In contrast, more than half of the remaining labels occur in less than 3% of the training set. Notably, the rarest class (label 14) appears only 251 times, accounting for just 0.5% of the data. Such a skewed distribution poses a challenge for multi-label learning, as frequent classes can dominate training loss while rare classes suffer from insufficient supervision.

To address this label skew, we apply a mixed resampling strategy to the training set after stratified 80%/20% train/validation split. Specifically, for training set, we perform oversampling on minority classes: for each label with fewer than 800 positive instances in the training set, we randomly replicate samples with that label until the class frequency reaches 800. We avoid downsampling frequent labels to preserve data diversity. This resampling strategy improves label coverage, mitigates class imbalance, and has been shown to improve generalisation in multi-label settings (Charte et al., 2015; Buda et al., 2018).

Following resampling, we employ aggressive data augmentation for the image modality. The training images are passed through a pipeline consisting of random rotation ($\pm 30^\circ$), random resized crop (scale $\in [0.7, 1.0]$), and 50% horizontal flipping. These transformations are designed to expand the effective training set and improve model robustness to spatial variation. The decision to use strong augmentation is motivated by recent findings in ViT literature (Dosovitskiy et al., 2021; Touvron et al., 2021), where larger receptive fields and attention-driven representations benefit from greater diversity in training views. Moreover, since our training set includes duplicated oversampled examples, strong augmentation reduces overfitting by introducing visual variability into repeated instances (Kang et al., 2020).

The validation set is preprocessed with center cropping after resizing to 400×400 , producing a consistent 384×384 view.

For the textual modality, we use the tokenizer associated with the pretrained BERT-Mini model. Captions are tokenized into WordPiece tokens and padded or truncated to a maximum length of 32. Each caption is encoded into a pair of token ID sequences and corresponding attention masks, both of shape [B, 32], and fed into the text encoder during training.

This preprocessing pipeline ensures a balanced, diverse, and standardized input space across modalities, enabling MiniViTex to effectively learn cross-modal representations in spite of data sparsity and imbalance.

3.8. Training Details

All models are trained for 10 epochs using a batch size of 32. For each run, we save the checkpoint that achieves the lowest validation loss, and use it for final evaluation on the validation set. During training, we apply mixed-precision optimization with automatic gradient scaling (using PyTorch AMP) to reduce memory footprint and accelerate training. All experiments are conducted using two NVIDIA L40S GPUs, enabling efficient parallelization and reduced wall-clock training time across the multimodal transformer models.

We adopt the AdamW optimizer with a differentiated parameter group strategy to accommodate the heterogeneity between pretrained and randomly initialised components:

- Visual and textual encoders are pretrained, and thus are fine-tuned using a low learning rate $\eta = 3 \times 10^{-5}$ and mild regularization ($\lambda = 0.02$);
- The remaining components include fusion transformer and classification head, are randomly initialized and trained with a larger learning rate $\eta = 10^{-4}$ and aggressive weight decay $\lambda = 0.1$, which helps regularise the model and prevent overfitting, particularly in transformer modules where dropout is used minimally.

Table 1. Performance and efficiency comparison of MiniViTex and image-only baselines. All metrics are reported on the best validation checkpoint.

Model	Epoch	Runtime (s)	Train mAP	Train F1	Val mAP	Val F1	FP16 Size (MB)	GFLOPs
MiniViTex_EfficientNetV2S	8	596.9	76.3%	92.1%	**66.8%**	**88.6%**	65.96	8.37
MiniViTex_ResNet50	9	629.2	**78.2%**	**92.8%**	63.8%	87.6%	73.21	12.02
MiniViTex_DeiTS	7	**277.9**	73.2%	90.7%	63.5%	86.9%	68.38	9.18
EfficientNetV2S + Attn	9	647.1	69.8%	89.3%	61.2%	86.3%	**41.36**	8.37
ResNet50 + Attn	8	520.8	71.7%	90.0%	56.5%	84.1%	48.61	12.01
DeiTS	9	310.3	66.5%	88.4%	56.3%	84.5%	43.79	**7.94**

We use a cosine learning rate schedule with linear warm-up. Let s be the current global step, S_{warm} the number of warm-up steps (10% of the training steps), and S_{total} the total number of training steps. The learning rate is scaled as:

$$\eta(s) = \begin{cases} \eta_0 \cdot \frac{s}{S_{\text{warm}}}, & \text{if } s < S_{\text{warm}} \\ \eta_0 \cdot 0.5\left(1 + \cos\left(\frac{\pi(s - S_{\text{warm}})}{S_{\text{total}} - S_{\text{warm}}}\right)\right), & \text{otherwise} \end{cases},$$

where η_0 is the base learning rate for each parameter group.

3.9. Evaluation Metrics

We evaluate models using mean Average Precision (mAP) and sample-wise F1 score (mean F1):

Mean Average Precision (mAP): Measures ranking quality across all labels. For each class, we compute Average Precision (AP) and average across classes:

$$\text{mAP} = \frac{1}{C}\sum_{j=1}^{C} \text{AP}_j.$$

Mean F1: Defined by sklearn as `f1_score(y_true, y_pred, average='samples')`. For each sample i, with predicted label set $\hat{Y}_i$ and true label set Y_i, the sample-level F1 is:

$$\text{F1}_i = \frac{2 \cdot |\hat{Y}_i \cap Y_i|}{|\hat{Y}_i| + |Y_i|}, \quad \text{Mean F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i.$$

This metric emphasizes correct prediction overlap on a per-instance basis, which is more sensitive to multi-label consistency than class-wise F1.

4. Experiments and Results

We conduct ablation experiments to evaluate the effectiveness of the proposed MiniViTex architecture under different image encoder configurations. For each encoder type EfficientNetV2-S, ResNet-50, and DeiT-S we train:

1. A full MiniViTex model with multimodal input (image + caption);
2. A corresponding image-only baseline using the same visual backbone but without caption input, where the model uses only the visual pseudo-CLS token for classification.

This design allows us to isolate the impact of text input and multimodal fusion across backbone choices.

All models are trained for up to 10 epochs with early stopping on validation loss. Performance is evaluated using mAP and mean F1 score (sample-wise) on both training and validation sets. From results Table 1, several key observations emerge:

Impact of Text Modality: The integration of textual information consistently enhances model performance across all evaluated backbones, confirming the value of our proposed multimodal fusion approach. Specifically, MiniViTex with multimodal fusion outperforms its corresponding image-only models:

- EfficientNetV2-S: mAP improve from 0.612 to 0.668 (+5.6%), mF1 improve from 0.863 to 0.886 (+2.4%).
- ResNet-50: mAP improve from from 0.565 to 0.638 (+7.3%), mF1 improve from 0.841 to 0.876 (+3.5%).
- DeiT-S: mAP improve from from 0.563 to 0.635 (+7.2%), mF1 improve from 0.845 to 0.869 (+2.4%).

These improvements indicate that even brief captions significantly contribute to reducing false positives and negatives, directly enhancing the consistency of predictions per sample as evidenced by increased Mean F1. This aligns with previous findings demonstrating the strength of early multimodal fusion in cross-modal contexts (Kim et al., 2021; Chen et al., 2020).

Comparison of Visual Backbones We observe distinct performance patterns between different backbone architectures. In image-only settings, EfficientNetV2-S achieves the highest validation mAP (0.612) and Mean F1 (0.863), surpassing ResNet-50 and DeiT-S significantly. With multimodal fusion, EfficientNetV2-S-based MiniViTex remains superior, achieving the highest validation metrics overall (mAP = 0.668, Mean F1 = 0.886).

Interestingly, while ResNet-50 exhibits the strongest performance on the training set (Train mAP = 0.782, Mean F1 = 0.928), its validation metrics (mAP = 0.638, Mean F1 = 0.876) suggest a potential risk of overfitting due to its deeper and more complex structure. In contrast, DeiT-based models achieve competitive efficiency (lowest runtime) but generally lower accuracy, which can be attributed to the reduced inductive biases inherent in pure transformer-based vision architectures (Dosovitskiy et al., 2021; Touvron et al., 2021).

Computational Efficiency and Trade-offs In terms of training runtime, DeiT-S-based MiniViTex models are the fastest (approximately 278s), followed by EfficientNetV2-S ($\approx 597s$) and ResNet-50 ($\approx 629s$). EfficientNetV2-S strikes the best balance, providing high accuracy (mAP: 0.678, Mean F1: 0.895) without substantial runtime penalty. This makes it a suitable choice for real-world scenarios requiring balanced performance and computational cost.

In addition to the runtime, we further analyzed the memory footprint and inference cost of each model. The overall parameter size (FP16) ranged from 65.96 MB (MiniViTex_EfficientNetV2S) to 73.21 MB (MiniViTex_ResNet50), which suggests that MiniViTex maintains a compact structure in all configurations. It is worth noting that despite the use of both image and text modes, the overall parameter growth is not significant compared to the vision-only backbone.The FLOPs of multiply-add operations. need for a single images also shows that MiniViTex remains efficient.The EfficientNetV2-S variant requires only 8.37 GFLOPs, and even the heaviest ResNet-50 version requires only 12.02 Even the heaviest ResNet-50 version requires only 12.02 GFLOPs. 8.37 GFLOPs are also applied to the DeiT-S-based MiniViTex, a strong comparison with EfficientNetV2-S at the same model complexity. The modest increase in FLOPs compared to its image-only counterpart demonstrates that our multimodal fusion design delivers a sizable performance gain while imposing an almost negligible computational overhead.

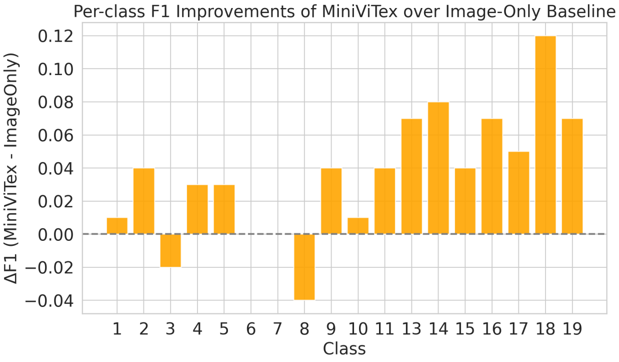
Overall, MiniViTex lives up to its name of “Mini” not only in size but also in compute—making it an excellent candidate for deployment in low-resource or real-time applications where both accuracy and efficiency are critical.

Summary of Multimodal Effectiveness Overall, the consistent improvements observed across both mAP and Mean F1 strongly support the effectiveness of MiniViTex's lightweight multimodal fusion strategy. By leveraging detailed cross-modal interactions between image patches and caption tokens at early stages, the model not only enhances overall ranking quality (mAP) but also achieves more coherent multi-label predictions at the individual sample level (Mean F1). This clearly demonstrates the advantage of token-level multimodal fusion over single-modality models.

Test Set Performance Due to the unavailability of test set labels, we submit the best-performing model, MiniViTex with EfficientNetV2-S backbone, to the official Kaggle leaderboard for final evaluation. The model achieves a mean F1 score of **94.08%** on the hidden test set, outperforming all baseline configurations and validating its generalisation ability beyond the validation set.

4.1. Per-Class Performance Analysis

Figure 1. Per-class F1 improvements of MiniViTex over the image-only EfficientNetV2-S baseline. MiniViTex significantly outperforms on animal-related and low-sample classes (e.g., Class 18), while performance slightly drops on transport-related categories (e.g., Class 8) due to textual dominance over background visual cues.



To further understand our model, we select the best performances variant MiniViTex EfficientNetV2-S compare to its image-only base line to conduct per-class performances. Table A.1 and Figure 1 reveal that MiniViTex outperforms the image-only baseline on 14 of 18 classes, with especially large gains on the long-tail categories. Below we highlight the most salient patterns.

Largest gain – Class 18 (+12% F1) Class 18 groups images whose main subject is an animal. Vision-only models struggle here because the visual appearance of animals is highly diverse (fur patterns, poses, environments), so a compact CNN often cannot learn discriminative filters for every species. MiniViTex, however, can rely on shallow lexical cues—common nouns such as cat, dog, horse—embedded in the caption. These cues guide the transformer to attend to

Table 2. Performance Comparison under Large Batch Size (128) and Scaled Learning Rate. All metrics are reported on the best validation checkpoint.

Model	Epoch	Runtime (s)	Train mAP	Train F1	Val mAP	Val F1
MiniViTex_EfficientNetV2-S	4	298.7	**75.9%**	**91.9%**	**66.4%**	**88.9%**
MiniViTex_ResNet-50	4	273.6	70.8%	89.9%	63.9%	87.9%
MiniViTex_DeiT-S	3	**112.5**	67.8%	88.3%	64.9%	87.2%
EffNetV2-S + Attn (Image Only)	4	279.6	69.7%	89.3%	60.2%	86.1%
ResNet-50 + Attn (Image Only)	4	254.5	62.9%	86.2%	52.7%	81.7%
DeiT-S (Image Only)	4	130.9	64.9%	87.5%	56.0%	84.1%

the correct image patches and compensate for limited visual prototypes, yielding the largest absolute improvement.

Moderate gains – Classes 13, 14, 16, 19 (+ 6 $\sim$ 7% F1) All four are low-support (¡ 100 samples) fine-grained semantic categories. In such scarce regimes, the added textual channel acts as weak supervision, providing label words or context tokens that stabilise training and curb overfitting.

Negative outlier – Class 8 (-4% F1) Class 8 corresponds to transport-related scenes. The image-only baseline simply fires whenever it detects a prominent vehicle. Captions, however, often focus on the salient foreground (e.g. "family picnic") and ignore background vehicles. In MiniViTex the caption tokens thus dominate early fusion, pushing transport-related patches down-weighted in self-attention and causing missed detections. Future work could mitigate this by (i) quality-filtering captions, or (ii) adding a late-fusion residual branch to preserve purely visual evidence.

Long-tail benefit overall Averaged over the eight classes with $\leq$ 300 training samples, MiniViTex with EfficientNetV2-S achieves $+6.5\%$ $\Delta F1$, versus $+1.7\%$ on high-frequency classes, confirming that text cues are most valuable where visual data are sparse.

In summary, the per-class breakdown corroborates our central claim: early multimodal fusion is particularly beneficial for semantically rich but visually heterogeneous or low-resource categories, while it may hurt when captions omit critical background objects.

4.2. Hyperparameter Analysis: Effects of Large Batch Size Training

Motivated by recent findings from OpenCLIP (Ilharco et al., 2021), suggesting that larger batch sizes may enhance the stability and performance of transformer-based models, we further explore the impact of increasing batch size from 32 to 128. Correspondingly, we scale the learning rate linearly by a factor of four, following the linear scaling rule (Goyal et al., 2018). Table 2 summaries the comparative results between the large batch (128) and the original baseline (32) setups across all MiniViTex configurations.

Training and Validation Performance A consistent observation across models is the reduction in training mean Average Precision (mAP) when increasing the batch size from 32 to 128. For instance, MiniViTex with EfficientNetV2-S shows a slight drop in training mAP from 76.3% to 75.9%, while more substantial decreases are observed for ResNet-50 (from 78.2% to 70.8%) and DeiT-S (from 73.2% to 67.8%).

Such a reduction suggests a decline in the model's ability to closely fit the training data at higher batch sizes. This phenomenon aligns with prior studies indicating that large batch sizes reduce gradient noise and stochasticity, potentially limiting the model's exploration of the parameter space during training (Keskar et al., 2017; Masters & Luschi, 2018).

However, this apparent underfitting effect compared to small-batch training does not negatively impact validation performance. On the contrary, we observe that validation metrics remain stable or even slightly improve. For example, MiniViTex with EfficientNetV2-S maintains similar validation mAP (66.8% → 66.4%) while slightly improving in mean F1 (88.6% → 88.9%). MiniViTex with ResNet-50 exhibits comparable stability (mAP: 63.8% → 63.9%, F1: 87.6% → 87.9%). Notably, the DeiT-S variant achieves a meaningful gain in both validation mAP (63.5% → 64.9%) and mean F1 (86.9% → 87.2%).

This counterintuitive improvement in validation metrics, despite reduced training metrics, suggests that larger batch sizes effectively regularise the model training. Specifically, the observed drop in training mAP accompanied by stable or improved validation performance indicates that models trained with smaller batches (batch size 32) might be overfitting, capturing noise or overly specific features of individual training examples. Larger batch sizes, by contrast, tend to produce smoother gradient updates, encouraging the model to learn more generalisable representations and reducing the risk of memorisation of the training set (Smith et al., 2018).

Training Stability and Efficiency Another critical observation is the enhanced stability during training afforded by larger batch sizes. Reduced stochasticity from increased batch size provides more stable gradient estimates, enabling the use of higher learning rates without the instability typically associated with aggressive training strategies. This advantage translates into accelerated convergence—achieving comparable or superior validation results with fewer epochs and substantially reduced computational runtime.

These results strongly suggest that the large batch regime not only mitigates overfitting but also enhances training efficiency. Such efficiency gains are particularly beneficial in scenarios involving limited computational resources or rapid experimentation cycles.

Practical Training Strategies In practical training scenarios, this analysis highlights a clear trade-off and balance that practitioners must navigate: smaller batches potentially maximise raw fitting performance but risk overfitting and instability, whereas larger batches provide stable gradients, regularise training effectively, and promote better generalisation. Thus, the choice of batch size should be guided by the model's sensitivity to overfitting, available computational resources, and the specific balance desired between training efficiency and optimal fit to the training data.

Future work may involve more fine-grained exploration of intermediate batch sizes or dynamic adjustment of batch size during training, potentially combining the stability and efficiency benefits of large batches with the fitting power of smaller batches.

5. Discussion

Our experiments show that MiniViTex, a unified, lightweight vision-and-text Transformer—achieves state-of-the-art accuracy on the Multi-Label Image-Caption Classification benchmark. Below we analyse the main empirical findings and their implications.

Text modality as a disambiguator Across all three visual backbones, injecting even very short captions improves validation mAP by $6 - 8$ *pp* and mean F1 by ≈ 3 *pp* compared with image-only counterparts. These gains confirm that the textual stream supplies high-level semantics (objects, scene type, sentiment) that help the network resolve visually ambiguous patterns, thereby lowering both false-positive and false-negative rates. Early token-level fusion inside the Transformer is crucial, because it allows fine-grained cross-attention between each caption token and every image patch, rather than relying on late feature concatenation.

Classification-only objective and a built-in path to alignment Unlike CLIP-style dual encoders, MiniViTex is trained solely on binary cross-entropy classification loss. This choice removes the heavy cost of hard-negative mining and yields faster convergence on our modest-sized (≈ 30 k sample) dataset. Yet the architecture retains separate $[\mathrm{CLS}]_{\mathrm{img}}$ and $[\mathrm{CLS}]_{\mathrm{txt}}$ tokens, both projected into a common 384-d space. This design keeps the door open to:

- **Contrastive fine-tuning.** A cosine-similarity (InfoNCE) term can be added between the two CLS vectors without introducing new parameters.
- **Multi-task learning.** Alignment and classification losses can be jointly optimised, balancing "what is in the picture" with "why this caption matches".
- **Zero-/few-shot capabilities.** Shared embeddings would enable prompt-based classification and cross-modal retrieval once alignment is added.

At present, however, the raw CLS embeddings remain weakly coupled; MiniViTex cannot yet support zero-shot transfer.

Backbone trade-offs The three tested vision encoders outline a clear Pareto frontier:

Table 3. Comparison of different visual backbones in MiniViTex. †Normalized to EfficientNetV2-S runtime.

Backbone	Val. mF1	Inference Time†	Params
EfficientNetV2-S	**88.9%**	1.00× (baseline)	41M
ResNet-50	87.9%	1.43×	55M
DeiT-Small	87.2%	**0.9×**	35M

EfficientNetV2-S offers the best accuracy but slightly longer inference time; ResNet-50 attains the second performance but a sharp increase in inference time and complexity; DeiT-S is fastest but slightly less accurate. Practitioners can therefore select a backbone along the speed–size–accuracy triangle according to deployment constraints.

Large-batch regularisation Increasing the batch size from 32 to 128 halves wall-clock training time and improves validation metrics despite lowering training-set precision. The smoother gradient estimate acts as an implicit regulariser, encouraging more generalisable minima. Tuning batch size to the available memory is therefore an efficient way to trade compute for quality.

6. Conclusion

We introduced MiniViTex, a compact multimodal Transformer that fuses EfficientNet/ResNet/DeiT visual features with a BERT-Mini caption encoder through early token concatenation. Without any explicit image–text alignment loss, MiniViTex reaches 66.8 % mAP and 88.9 % mean-F1, exceeding image-only baselines by a wide margin. The architecture deliberately preserves modality-specific CLS tokens, giving practitioners a seamless upgrade path to CLIP-style contrastive learning and zero-shot inference—without rewriting the model or re-initialising weights.

Future research will (i) add alignment and grounding losses to unlock retrieval and explainability, (ii) explore adaptive batch-size curricula, and (iii) compress the model further for edge deployment. Taken together, these results establish MiniViTex as a robust, efficient, and extensible template for real-world vision–language applications.

References

Bhargava, P., Drozd, A., and Rogers, A. Generalization in nli: Ways (not) to go beyond simple heuristics, 2021.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.07.011. URL https://www.sciencedirect.com/science/article/pii/S0893608018302107.

Charte, F., Rivera, A. J., del Jesus, M. J., and Herrera, F. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2014.08.091. URL https://www.sciencedirect.com/science/article/pii/S0925231215004269. Recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems Progress in Intelligent Systems Mining Humanistic Data.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, pp. 104–120, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58577-8.

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.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018. URL https://arxiv.org/abs/1706.02677.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hugging Face. DeiT — Transformers documentation. https://huggingface.co/docs/transformers/model_doc/deit, n.d. Accessed: 2025-05-23.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition, 2020. URL https://arxiv.org/abs/1910.09217.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima, 2017. URL https://arxiv.org/abs/1609.04836.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. URL https://arxiv.org/abs/2102.03334.

Kobayashi, S., Akram, Y., and Oswald, J. V. Weight decay induces low-rank attention layers, 2024. URL https://arxiv.org/abs/2410.23819.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language, 2019. URL https://arxiv.org/abs/1908.03557.

Lin, J. Multi-label classification competition 2025. https://kaggle.com/competitions/multi-label-classification-competition-2025, 2025. Kaggle.

Liu, L., Liu, X., Gao, J., Chen, W., and Han, J. Understanding the difficulty of training transformers, 2023. URL https://arxiv.org/abs/2004.08249.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. URL https://arxiv.org/abs/1908.02265.

Masters, D. and Luschi, C. Revisiting small batch training for deep neural networks, 2018. URL https://arxiv.org/abs/1804.07612.

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.

Smith, S. L., Kindermans, P.-J., Ying, C., and Le, Q. V. Don't decay the learning rate, increase the batch size, 2018. URL https://arxiv.org/abs/1711.00489.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7242–7252, 2021. doi: 10.1109/ICCV48922.2021.00717.

Tan, H. and Bansal, M. Lxmert: Learning cross-modality encoder representations from transformers, 2019. URL https://arxiv.org/abs/1908.07490.

Tan, M. and Le, Q. V. Efficientnetv2: Smaller models and faster training, 2021. URL https://arxiv.org/abs/2104.00298.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers distillation through attention, 2021. URL https://arxiv.org/abs/2012.12877.

Turc, I., Chang, M., Lee, K., and Toutanova, K. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR*, abs/1908.08962, 2019. URL http://arxiv.org/abs/1908.08962.

A. Per-class Evaluation Details

Table A.1. Per-class evaluation metrics for MiniViTex (left) and image-only baseline (right). Best F1-score per class is in bold.

Class	MiniViTex			Image-only			Support	ΔF1
	Precision	Recall	F1	Precision	Recall	F1		
1	0.96	0.94	0.95	0.96	0.92	0.94	4561	+0.01
2	0.91	0.50	0.65	0.81	0.49	0.61	244	+0.04
3	0.80	0.52	0.63	0.77	0.57	0.65	865	–0.02
4	0.98	0.78	0.87	0.94	0.76	0.84	268	+0.03
5	0.99	0.97	0.98	0.97	0.93	0.95	219	+0.03
6	0.90	0.75	0.82	0.89	0.76	0.82	286	±0.00
7	0.98	0.87	0.92	0.96	0.89	0.92	245	±0.00
8	0.78	0.47	0.59	0.76	0.54	0.63	463	–0.04
9	0.94	0.72	0.81	0.86	0.71	0.77	201	+0.04
10	0.77	0.66	0.71	0.78	0.64	0.70	281	+0.01
11	0.84	0.68	0.75	0.79	0.65	0.71	96	+0.04
13	0.88	0.55	0.68	0.71	0.53	0.61	89	+0.07
14	0.95	0.73	0.82	0.84	0.67	0.74	48	+0.08
15	0.75	0.50	0.60	0.71	0.46	0.56	407	+0.04
16	0.92	0.66	0.77	0.79	0.63	0.70	237	+0.07
17	0.92	0.91	0.92	0.89	0.86	0.87	300	+0.05
18	0.95	0.80	0.87	0.87	0.66	0.75	295	+0.12
19	0.99	0.89	0.94	0.91	0.83	0.87	199	+0.07

B. Code, Model, and Reproducibility Instructions

B.1. Access Link

Google Drive:

https://drive.google.com/drive/folders/1uTqBoqjR9w49ilCG_RpYu541IpmzAaS9?usp=sharing

B.2. File Structure

```
experiment.ipynb            # Train all MiniViTex variants and image-only baselines
test.ipynb                  # Load best model and predict test labels
Predicted_labels.txt        # Output file containing test predictions
best_model/                 # Best-performing MiniViTex (EfficientNetV2-S) checkpoint
experiments/                # Training logs
image_only/                 # Output from image-only (EfficientNetV2-S) model
```

B.3. Software Environment

OS: Ubuntu 22.04, **Python:** 3.11, **PyTorch:** 2.4.0, **cuda:** 12.4.1, **Mixed Precision:** Enabled via `torch.cuda.amp`

B.4. Hardware Environment

GPU: 2 × NVIDIA L40S (48GB VRAM each), **Training Mode:** Fully parallelized using PyTorch `DataParallel`

B.5. Training Instructions

Open and run `experiment.ipynb`. It will:

- Train all MiniViTex variants (EfficientNetV2-S, ResNet-50, DeiT-S)
- Train corresponding image-only baselines
- Save the best validation checkpoint
- Log per-epoch performance in CSV:

```
epoch,runtime_sec,train_loss,train_map,train_f1_macro,train_f1_micro,train_f1_mean,
val_loss,val_map,val_f1_macro,val_f1_micro,val_f1_mean
```

B.6. Inference Instructions

Open and run `test.ipynb`. It will:

- Load the best MiniViTex model (EfficientNetV2-S)
- Predict labels for the test set
- Save predictions to `Predicted_labels.txt`

No internet access or pretrained model downloads are required during inference, all needed model weight and tokenizer are in `best_model/` folder.