# Visually Informed Text Representations for Visual Contextual Classification of Arts

Raphaëlle Lemaire[1 [0009−0000−8323−8855]], Jérémie Pantin[1 [0009−0002−5082−6815]], Alexis Lechervy[1 [0000−0002−9441−0187]], Azamat Kaibaldiyev[1 [0009−0002−0425−0798]], Fabrice Maurel[1 [0000−0002−8644−2461]], Gaël Dias[1 [0000−0002−5840−1603]], and Youssef Chahir[1 [0000−0002−1417−5317]]

Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

**Abstract.** Artworks' textual descriptions often intertwine visual observations with external contextual knowledge. Automatically distinguishing between these two types of information is challenging for language models. We address the gap between unimodal language models, which lack visual grounding, and multimodal models, which assume complete text–image alignment. Our approach learns visually-informed text representations by integrating visual knowledge into a BERT-based unimodal language model, considering partial text-image alignment. We train a text encoder to attend to visual features via cross-attention, then extract the adapted encoder for text-only inference. This representation learning strategy enables the model to internalize visual grounding without requiring images at deployment. Experiments on Artpedia and AQUA datasets show consistent gains of our model (VBB) over a wide range of unimodal and multimodal baselines.

**Keywords:** Multimodal learning · Art analysis · Text classification.

## 1 Introduction

Textual descriptions of artworks typically consider two kinds of information: *visual* and *contextual*. Visual information refers to elements directly observable in the image, while contextual information focuses on external knowledge about the artwork (life of the painter for example). Separating these strands is non-trivial due to the breadth and subtlety of art discourse, where implicit references, symbolism, and style conventions are common [7]. In this paper, we formalise Visual–Contextual Text Classification (VCTC), a binary classification task that aims to determine whether a textual description associated with an artwork is visually grounded or contextual. VCTC is relevant for a variety of downstream applications in AI-for-Art, including visual question answering [13, 6, 33], text-image retrieval [34], and artwork description generation [4]. Despite this relevance, VCTC has received limited attention as a standalone problem. In particular, its properties such as semantic ambiguity, partial visual groundability, and the absence of token-level supervision, have not been systematically addressed.

From a representation learning perspective, VCTC poses a specific challenge. Unimodal language models (LMs), such as BERT [9], rely exclusively on textual co-occurrences and therefore struggle to identify linguistic patterns whose interpretation depends on visual evidence [3]. Multimodal models, such as CLIP [30], learn joint image-text representations through global alignment objectives, but they are primarily designed for fully grounded multimodal tasks. As a result, they typically require images at inference time, and their text encoders are not optimised for fine-grained, text-only classification [29, 28, 37, 12].

We argue that VCTC benefits from fused representations learned during training, while still requiring text-only inference. Rather than enforcing full image-text alignment, the goal is to specialize textual representations so that they become sensitive to visual groundability cues. To this end, we propose a training framework in which a unimodal text encoder is temporarily coupled with an image encoder via cross-attention. During training, this fusion mechanism allows textual tokens to selectively attend to visual regions, encouraging the text encoder to internalise which linguistic elements correspond to observable visual evidence. At inference time, the image encoder and fusion module are discarded, yielding an efficient text-only model whose representations retain visual awareness. We evaluate our approach on the Artpedia [34] and AQUA [13] datasets, which expose complementary aspects of the visual-contextual divide in art descriptions. Across both datasets, we show that text encoders trained with visually informed fusion consistently outperform strong unimodal baselines and off-the-shelf multimodal encoders when applied to text-only VCTC.

Three contributions are presented in this work. First, we consider VCTC as a primary task for analyzing art descriptions, clarifying task definitions and challenges (ambiguity, symbolism, and partial groundability). Secondly, we introduce a cross-attentional training strategy that learns visually informed text representations through multimodal fusion, while preserving text-only inference. Finally, we present the first comprehensive study of VCTC as a leading problem on both Artpedia and AQUA. Our results demonstrate that explicit visual-informed training yields robust gains over both unimodal LMs and off-the-shelf multimodal encoders.

## 2 Related work

The classification of art descriptions into visual and contextual categories (VCTC) sits at the intersection of several research domains. While the literature recognizes that art descriptions naturally interleave visual observations with contextual knowledge [7, 4], and some works leverage this distinction in broader pipelines such as VCTC's integration of context-aware VQA [13, 5] and enhanced retrieval beyond purely visual captions [34] treating this separation as a primary classification problem remains uncommon. Existing models typically optimize either for fully grounded image-text matching or for generic text categorization, leaving VCTC's core challenges under-addressed.

*Text-only approaches* Transformer-based language models [36], such as BERT [9] have become standard for text classification through bidirectional self-attention and strong transfer capabilities [11, 14, 1]. However, these models lack visual grounding, limiting their ability to judge whether statements are visually verifiable. Recent efforts attempt to inject visual priors into language models without changing their inference modality: some transfer CLIP-derived signals into BERT [3], others use distillation approaches with vision-language teachers like OSCAR [22], and some directly pretrain with visually linked corpora [40, 2]. Despite these advances, such approaches predominantly probe surface-level visual attributes (color, shape) and do not target the partial groundability and semantic ambiguity characteristic of art descriptions.

*Multimodal approaches* Joint image-text models have proven effective at capturing rich semantics through modality alignment or fusion [38]. Contrastive methods like CLIP [30], ALBEF [21], and their text-only variants [29, 41] optimize global alignment in shared embedding spaces, making them well-suited for retrieval but too coarse for sentence-level groundability decisions. Fusion-based transformers such as UNITER [8], LXMERT [35], ViLT [17], and BLIP [19, 18, 39] use cross-attention or single-stream mechanisms, while instruction-tuned VLMs like LLaVA [23] tackle diverse multimodal scenarios. Despite their success on VQA and captioning, these models require images at inference time and are trained with alignment objectives.

*Text-only inference in art* In classification settings where only text is available at test time, incorporating images can be unnecessary or even counterproductive. For instance, some studies on encyclopedic categorization report mixed gains from adding images [26, 27]. In art analysis, downstream tasks (e.g. VQA, retrieval, description) often entangle grounding with recognition, obscuring the distinct challenge of separating visual from contextual information. These observations imply that the VCTC task requires text-only inference, while still benefiting from visual priors. There remains a need for methods that transfer visual knowledge into unimodal text representations during training, while preserving text-only inference and explicit optimization for partial groundability Such methods would complement purely textual baselines [9] and fully multimodal pretraining ones [30, 21, 8, 35].

## 3 Visual-contextual text classification task

We introduce the Visual-Contextual Text Classification (VCTC) task as a binary classification problem at the document level, with an optional token-level analysis. Let $I = \{I_i \mid i = 1, \ldots, N_{\text{img}}\}$ be the set of artwork images and $D = \{D_j \mid j = 1, \ldots, N_{\text{doc}}\}$ the set of textual descriptions. Each document $D_j$ is a sequence of tokens $D_j = (t_{j,1}, \ldots, t_{j,L_j})$, where $L_j \in \mathbb{N}$ is the number of tokens. Typically, $L_j$ refers to a fixed-length padding. Each document is associated to exactly one image with a mapping $f_{\text{img}} : \{1, \ldots, N_{\text{doc}}\} \to \{1, \ldots, N_{\text{img}}\}$.



**Art Description (Wikipedia)**

***[contextual]*** The Portrait of Luca Paciolo, ***[visual]*** a friar and mathematician with a table filled with geometrical tools: slate, chalk, compass, dodecahedron modelis ***[contextual]*** a painting attributed to the Italian Renaissance artist Jacopo de Barbari, dating to around 1500 and housed in the Capodimonte Museum, Naples, southern Italy.

**Textual and visual questions (AQUA)**

***[contextual]***
Who was also called Luca Di Borgo after his birthplace? Fra Luca Bartolomeo de Pacioli
What is Pacioli demonstrating by Euclid? A theorem

***[visual]***
What is the table made of? Wood
What is the man holding? Court

**Textual and visual documents (Artpedia)**

***[contextual]*** The Portrait of Luca Pacioli is a painting attributed to the Italian Renaissance artist Jacopo de Barbari, dating to around 1500 and housed in the Capodimonte Museum.

***[visual]*** The painting portrays the friar and mathematician with a table filled with geometrical tools: slate, chalk, compass and a dodecahedron model.

**Fig. 1.** Distinction between visual and contextual information in descriptions, question answering, and document classification.

As images may be paired with multiple documents of different labels, supervision is applied exclusively at the document level. Label space is $\mathcal{Y} = \{0, 1\}$, where a document (optionally a token) is visual (1) if its content is verifiable from the paired image $I_{f_{\text{img}}(j)}$ alone, and contextual (0) otherwise. In particular, a contextual label may refer to external or art-historical knowledge beyond $I_{f_{\text{img}}(j)}$. Figure 1 illustrates this situation.

*Document vs. token labels.* In order to understand the classification process in depth, it can be useful to represent text mentions as a set of individual tokens. As such, it is possible to evaluate the contribution of each token individually to the document classification task. As a consequence, we may define token labels $z_{j,k} \in \{0, 1\}$ indicating whether token $t_{j,k}$ expresses visual content (1) or contextual content (0). In practice, datasets typically provide $y_j$ only at document-level and token labels $z_{j,k}$ are unobserved. So, when we analyze token-level behavior, we use weak supervision by propagating the document label $y_j$ to all tokens contained in $D_j$ such that $z_{j,k} = y_j$.

*Training and inference.* The VCTC training objective is to estimate a text classifier that predicts $p(y_j = 1 \mid D_j)$ and, optionally, token posteriors $p(z_{j,k} = 1 \mid D_j)$ for in-depth understanding. Crucially, inference is text-only, i.e. at test time, predictions must depend on $D_j$ without requiring access to $I_{f_{\text{img}}(j)}$. During training, paired images may be used to guide representation learning for encouraging sensitivity to image-groundable cues. So, the problem definition itself does not assume images at test time. This formulation isolates the VCTC classification objective to decide whether a textual unit is image-groundable. Thus, we

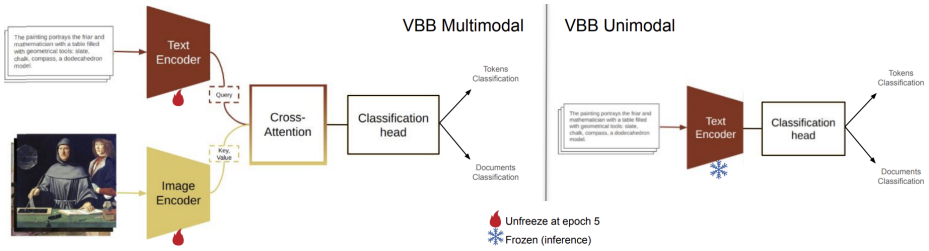


**Fig. 2.** Overview of $\text{VBB}_\text{m}$ (left) for token and document classification of the VCTC task. A text encoder (BERT) and an image encoder (ViT) extract representations from the input document and its corresponding image. A cross-attention module then allows each text token to query relevant visual regions by using textual embeddings as queries and image patches as keys and values. The fused representations are then passed through a classification head. On the right, a presentation of $\text{VBB}_\text{u}$ with the text encoder extracted from $\text{VBB}_\text{m}$ and the addition of a classification layer to predict token classes.

keep document-level evaluation and fine-grained token-level inspection within the same framework.

## 4 The VBB architecture

Our ViT-Boosted BERT model (VBB) for VCTC is illustrated in Figure 2. In particular, it combines multimodal classification for training, and visually grounded text-only representation at inference.

### 4.1 Encoders

For documents, we use a pretrained LM such as BERT [9] and cap sequences to $L_\text{max}$ tokens. With $\mathcal{D}_j \in \mathbb{R}^{L_{max} \times d_\text{T}}$ denote the hidden state, with $d_\text{T}$ being the BERT dimension, after the last Transformer layer. We compute $\mathcal{D}_j$ by processing all the tokens in $\left(D_j\right)$ with the pretrained LM. For images, we use a pretrained transformer-based model ViT [30]. After patchification, we obtain $P_{patch}$ patch embeddings and $\mathcal{I}_j \in \mathbb{R}^{P_{patch} \times d_\text{I}}$ is the hidden state with $d_\text{I}$ being the ViT dimension, with 2D positional encodings to preserve spatial layout. We compute $\mathcal{I}_j$ by passing the image $I_{f_\text{img}(j)}$ to ViT.

### 4.2 Cross-attention fusion

We project both modalities to a common model dimension $d$. Thus, we note $\hat{\mathcal{D}}_j = \mathcal{D}_j W^\text{D}$ with $W^\text{D} \in \mathbb{R}^{d_\text{T} \times d}$, and $\hat{\mathcal{I}}_j = \mathcal{I}_j W^\text{I}$ with $W^\text{I} \in \mathbb{R}^{d_\text{I} \times d}$. Multi-head cross-attention (MHA) uses token queries and image keys/values:

$$Q = \hat{\mathcal{D}}_j W_Q, K = \hat{\mathcal{I}}_j W_K, V = \hat{\mathcal{I}}_j W_V. \quad (1)$$

The fused token states are defined as $\tilde{\mathcal{D}}_j$, where LN is a residual layer norm:

$$\tilde{\mathcal{D}}_j = \mathrm{LN}\big(\hat{\mathcal{D}}_j + \mathrm{MHA}(Q, K, V)\big). \tag{2}$$

The fused representations are then passed through a number of H self-attention layers to integrate visual context across the sequence, each layer being optionally followed by a feed-forward block and a residual layer norm.

### 4.3 Token-level representation

Each fused token $\tilde{d}_{j,k}$ is transferred to a one-layer multilayer perceptron (MLP) with a ReLU activation function to produce a logit $\ell_{j,k}$, where $W_1$ is the weight matrix of the first layer, $w_2$ the weight vectors of the output layer, $b_1$ the bias vector of the first layer and $b_2$ the bias vector of the output layer:

$$\ell_{j,k} = w_2^\top \ \mathrm{ReLU}(W_1 \tilde{d}_{j,k} + b_1) + b_2. \tag{3}$$

Note that at inference, we can deduce the probability that token $t_{j,k}$ is visual or contextual by calculating $p_{j,k} = \sigma(\ell_{j,k})$, where $\sigma(\cdot)$ is the sigmoid function.

### 4.4 Document-level classification

To obtain the document-level classification, for one document $\mathcal{D}_j$, the logits of each token $\ell_{j,k}$ are aggregated to create a single document logit $\ell_j$:

$$\ell_j = \sum_{k=1}^{L_j} \ell_{j,k} \frac{\exp(\ell_{j,k})}{\sum_{m=1}^{L_j} \exp(\ell_{j,m})}. \tag{4}$$

The softmax aggregation assigns higher weights to tokens with stronger confidence, emphasizing highly informative visual tokens while down-weighting contextual ones, in accordance with the definition of visual documents. We use this approach to ensure that the most relevant and informative tokens contribute the most to the final document classification. We deduce the probability that document $\mathcal{D}_j$ is visual or contextual with $p_j = \sigma(\ell_j)$.

### 4.5 Training objective and class imbalance

Since our batch sizes are small (32), the class distribution is batch-dependent. To address this desequilibre, we implement a weighted version of the binary cross-entropy loss [10] which depends on the positive and negative example proposition present in the batches. At each training iteration, we compute the number of positive examples $E_p$ (visual) and negative examples $E_n$ (contextual) in the current batch with $E$ example. We define a class weight $\beta = \frac{E_n}{E_p}$ to emphasize the contribution of rare positive cases. The resulting loss function is:

$$\mathcal{L} = -\beta \sum_{j}^{E} [y_j \log(p_j) + (1 - y_j) \log(1 - p_j)] \tag{5}$$

where $E$ is the number of documents in the batch and $y_j \in \{0, 1\}$ is the corresponding binary label. The dynamic weight $\beta$ is recalculated for every batch.

### 4.6 Training and inference

*Training (multimodal).* Unless stated otherwise, we initialize images with the ViT encoder and texts with the BERT encoder, project the multimodal representation to a shared space $d$, and train the classification model end-to-end. Cross-attention couples gradients so the text encoder learns which linguistic cues are image-groundable.

*Inference (text-only).* At test time, we discard the image encoder and the cross-attention module, and run the adapted text encoder on $D_j$ alone to produce $p_{j,k}$ (and $p_j$ if the document head is used). This preserves deployment simplicity, while retaining benefits from multimodal co-training.

## 5 Experimental setup and VCTC results

### 5.1 Datasets

We evaluate our approach on two datasets that distinguish between visual and contextual textual information: Artpedia [34][1] and AQUA [13][2].

*Artpedia.* The Artpedia dataset consists of 2,930 paintings, each with textual descriptions manually annotated as either visual (describing perceivable elements in the image) or contextual (requiring external knowledge). In total, the dataset comprises 9,173 visual and 19,039 contextual documents.

*AQUA.* The AQUA dataset includes 21,384 artworks where images paired with question-answer pairs, annotated to indicate whether the answer requires visual information from the image or external knowledge. The same image may appear multiple times with different questions. The full dataset contains 32,345 visual and 47,503 contextual questions.

### 5.2 Learning setup

*Evaluation metrics.* For the VCTC task, we perform evaluation at two levels: document and token level. We use four metrics: micro accuracy ($\text{Acc}_{\text{micro}}$), macro accuracy ($\text{Acc}_{\text{macro}}$), F1 score, and Area Under ROC Curve (AUC).

*Models and baselines.* We evaluate the performance of our VBB model under unimodal mode ($\text{VBB}_{\text{u}}$), where only the text encoder is taken into account for inference, against several strong text-only baselines, including BERT and its variants like RoBERTa [24, 31], DeBERTa [16, 15] and DistilBERT [31]. We also include unimodal models extracted from multimodal architectures, e.g. the BERT text encoder from CLIP [30], and the BERT text encoder from BLIP [20]. For multimodal comparison, we use our VBB model under multimodal mode ($\text{VBB}_{\text{m}}$),

[1] http://aimagelab.ing.unimore.it/artpedia
[2] https://github.com/noagarcia/ArtVQA/tree/master/AQUA.

where both text and image encoders are taken into account for inference, and benchmark it against models such as VIKING [13], CLIPText [29], LABCLIP [41], BLIP [20], BLIP-2 [18], BLIP-3 [39], LLaVA [23] and ViLT [17]. For all models, we use recommended settings from the corresponding authors and use their codes (when available). Regarding LLaVA, we explore two adaptation strategies. We added a classification head is appended to the frozen backbone and we used prompting. For the latter, we report the results from the most effective prompt configuration, following [32]. To ensure fairness, we report the mean and standard deviation over five independent runs for all tested configurations.

*VBB training details.* $\text{VBB}_\text{m}$ and $\text{VBB}_\text{u}$ are trained using AdamW [25], and early stopping on validation loss. Presented values are set with patience = 5 and mean±std is calculated over five different seeds. $\text{VBB}_\text{m}$ initializes BERT and ViT from public checkpoints, while cross-attention and classifiers are randomly initialized. On the first stage (frozen encoders during 4 epochs), we only train fusion and heads' weights ($\text{lr} = 1 \times 10^{-4}$, $\text{wd} = 1 \times 10^{-4}$). On the second stage (end-to-end starting after 4 epochs), all modules are unfrozen ($\text{lr} = 1 \times 10^{-5}$). Finally, $\text{VBB}_\text{u}$ reuses the text encoder from $\text{VBB}_\text{m}$ and adds a 1-layer MLP classifier. We warm up the head ($\text{lr} = 1 \times 10^{-6}$, $\text{wd} = 1 \times 10^{-5}$), then fine-tune the encoder ($\text{lr} = 5 \times 10^{-5}$, $\text{wd} = 1 \times 10^{-5}$). All text sequences are capped at $L_{\max} = 77$ tokens.

*Hyperparameters and implementation.* Hyperparameters are selected based on the validation set with a small log-scale grid and fixed seeds. The best validation configuration is used for test-time reporting. Exact grids, seeds, and per-dataset selections appear in the supplementary materials. All models are implemented in PyTorch and trained on a single NVIDIA A100 (80GB).

### 5.3 Document-level classification results

*Text-only models.* The top part of tables 1 and 2 illustrates overall results for pure text representations models, where only text is given as input at training and inference. On Artpedia, BERT achieves the strongest results among pure unimodal models (79.50% micro accuracy, 83.39% macro accuracy, 75.37% F1), outperforming DeBERTa models. DistilBERT obtains the highest AUC (92.77%), indicating better differentiation of classes despite lower accuracy. On AQUA, DeBERTa shows the best results across all metrics.

*Multimodal models.* The middle part of tables 1 and 2 illustrates overall results for the multimodal models that jointly process images and text during both training and inference for the VCTC classification task. On Artpedia, Prompt-CLIPText (Ensemble) excels at class balance (84.49% macro accuracy, 92.34% AUC ROC), while standard CLIPText achieves strong macro accuracy (84.20%). However, our $\text{VBB}_\text{m}$ CLIP BERT attains the highest micro accuracy (84.98%), while $\text{VBB}_\text{m}$ BERT achieves best macro accuracy (84.93%) and F1 score (78.18%). This indicating that global contrastive pretraining still yields

| | | Artpedia | | | |
|---|---|---|---|---|---|
| | | $\text{Acc}_{\text{micro}}$ | $\text{Acc}_{\text{macro}}$ | AUC | F1 score |
| Unim | DeBERTa V3 [15] | 69.28±0.14 | 74.54±0.12 | 89.93±0.26 | 66.24±0.11 |
| | DistilBERT [31] | 75.23±0.34 | 80.51±0.30 | **92.77±0.26** | 71.96±0.31 |
| | BERT [9] | **79.50±0.37** | **83.39±0.21** | 90.89±0.14 | **75.37 ± 0.28** |
| Multimodal | CLIPText [29] | 82.91±0.06 | 84.20±0.13 | 91.18±0.27 | 77.73±0.08 |
| | CLIPText (Ensemble) [29] | 80.59±0.08 | 82.65±0.04 | 92.14±0.04 | 74.98±0.58 |
| | Prompt-CLIPText [29] | 83.28±0.13 | 83.81±0.07 | 90.05±0.67 | 77.19±0.09 |
| | Prompt-CLIPText (Ensemble) [29] | 83.88±0.06 | 84.49±0.00 | **92.34±0.31** | 78.00±0.03 |
| | LABCLIP [41] | 79.97±0.62 | 83.00±0.39 | 92.02±0.42 | 75.72±0.44 |
| | BLIP [20] | 74.72±0.10 | 78.82±0.14 | 88.74±0.57 | 70.39±0.12 |
| | LLaVA [23] | 79.19±0.12 | 79.67±0.05 | 87.54±0.11 | 71.77±0.07 |
| | Prompt LLaVA [23] | 75.51±0.47 | 68.10±0.91 | - | 55.14±1.76 |
| | ViLT [17] | 78.06±0.15 | 80.90±0.11 | 90.62±0.10 | 72.75±0.11 |
| | $\text{VBB}_{\text{m}}$ DeBERTa V3 | 78.96±0.21 | 77.72±0.41 | 85.88±1.25 | 69.97±0.62 |
| | $\text{VBB}_{\text{m}}$ DistilBERT | 80.50±0.27 | 82.69± 0.15 | 90.79± 0.43 | 75.23±0.22 |
| | $\text{VBB}_{\text{m}}$ BERT | 83.61±0.10 | **84.93±0.46** | 91.31±0.25 | **78.18±0.71** |
| | $\text{VBB}_{\text{m}}$ CLIP BERT | **84.98±0.25** | 83.66±0.34 | 91.02±0.31 | 77.83±0.36 |
| Distilled | CLIP BERT [30] | 71.81±0.24 | 77.37±0.23 | 88.16±0.39 | 68.75±0.17 |
| | BLIP BERT [20] | 73.02±0.18 | 76.82±0.12 | 88.48±0.21 | 68.37±0.13 |
| | $\text{VBB}_{\text{u}}$ DeBERTa V3 | 76.59±0.21 | 78.63±0.16 | 86.95±0.37 | 70.51±0.18 |
| | $\text{VBB}_{\text{u}}$ DistilBERT | **84.10±0.14** | 84.41±0.67 | **91.65±0.54** | **78.05±0.12** |
| | $\text{VBB}_{\text{u}}$ BERT | 83.25±0.17 | 84.16±0.56 | 91.49±0.65 | 77.46±0.48 |
| | $\text{VBB}_{\text{u}}$ CLIP BERT | 81.94±0.41 | **84.46±0.45** | 91.24±0.32 | 77.09±0.51 |

**Table 1.** Results for multimodal, unimodal (unim) and distilled unimodal models in VCTC on the Artpedia datasets, measured using micro accuracy ($\text{Acc}_{\text{micro}}$), macro accuracy ($\text{Acc}_{\text{macro}}$), Area Under the ROC Curve (AUC) and the F1 score. For all models, performance is reported as the mean ± standard deviation over five independent runs. The best result for each metric is highlighted in bold and the second best result is underlined.

a small edge on class balance and calibration. On AQUA, the VCTC task appears less challenging, with multiple models achieving near-ceiling performance. $\text{VBB}_{\text{m}}$ CLIP BERT reaches 99.88% micro accuracy and 99.77% F1, matching the specialized VIKING baseline while maintaining architectural flexibility.

*Distilled models.* The last part of tables 1 and 2 shows results for distilled models, where image and text were used at training and only text at inference. On Artpedia, all $\text{VBB}_{\text{u}}$ variants outperform other distilled models. $\text{VBB}_{\text{u}}$ DistilBERT achieves 84.10% micro accuracy versus 73.02% for BLIP BERT and 71.81% for CLIP BERT. This gap indicates that our two-stage process (co-training with cross-attention, then extraction and fine-tuning) transfers visual knowledge more effectively than contrastive or generative pretraining approaches. On Aqua, we see the same behavior as multimodals models with $\text{VBB}_{\text{u}}$ outperforming all others models and also solved the AQUA dataset with 99% on all metrics except for DeBERTa V3.

*Cross-category comparison.* Closely matching between $\text{VBB}_{\text{u}}$ and $\text{VBB}_{\text{m}}$ variants demonstrate that visual knowledge can be internalized into text encoders through multimodal co-training, then retained during text-only inference. On Artpedia, $\text{VBB}_{\text{u}}$ DistilBERT not only surpasses all text-only models but also exceeds several multimodal baselines, including BLIP (74.72%), LLaVA (79.19%), and ViLT (78.06%). On class-sensitive metrics (macro accuracy, AUC), distilled

| | | Aqua | | | |
|---|---|---|---|---|---|
| | | $\text{Acc}_{\text{micro}}$ | $\text{Acc}_{\text{macro}}$ | AUC | F1 score |
| Unim | DeBERTa [16] | **97.01±0.16** | **90.54±0.21** | **98.65±0.50** | **86.87±0.46** |
| | DeBERTa V3 [15] | 94.89±0.14 | 87.68±0.11 | 97.02±0.25 | 86.41±0.17 |
| | DistilBERT [31] | 95.35±0.27 | 90.48±0.29 | 98.32±0.22 | 86.39±0.17 |
| | BERT [9] | 84.90±0.09 | 89.76±0.08 | 92.14±0.07 | 77.37 ± 0.11 |
| Multimodal | VIKING classifier [13] | 99.52 ±0.01 | 99.22±0.19 | **99.93±0.05** | 99.28±0.17 |
| | CLIPText [29] | 96.33±0.03 | 82.81±0.11 | 99.16±0.43 | 80.27±0.18 |
| | CLIPText (Ensemble) [29] | 95.76±0.09 | 82.89±0.42 | 99.51±0.05 | 79.25±0.66 |
| | Prompt-CLIPText [29] | 94.47±0.03 | 74.61±0.18 | 96.51±0.41 | 66.0±0.31 |
| | Prompt-CLIPText (Ensemble) [29] | 94.84±0.04 | 75.48±0.20 | 96.40±0.09 | 67.46±0.36 |
| | LABCLIP [41] | 96.54±0.01 | 83.78±0.04 | 96.49±0.13 | 80.42±0.06 |
| | BLIP [20] | 90.23±0.13 | 88.84±0.26 | 94.86±0.39 | 84.47±0.23 |
| | ViLT [17] | 95.50±0.24 | 90.27±0.50 | 98.32±0.88 | 87.33±0.64 |
| | $\text{VBB}_{\text{m}}$ DeBERTa V3 | 98.88±0.06 | 88.07±0.09 | 98.29±0.06 | 91.86±0.17 |
| | $\text{VBB}_{\text{m}}$ DistilBERT | 95.62±0.02 | 91.54±0.03 | 98.54±0.04 | 90.75±0.04 |
| | $\text{VBB}_{\text{m}}$ BERT | 96.60±0.02 | 93.44±0.05 | 93.42±0.05 | 92.75±0.06 |
| | $\text{VBB}_{\text{m}}$ CLIP BERT | **99.88±0.06** | **99.85±0.07** | 99.52±0.05 | **99.77 ±0.12** |
| Dist | CLIP BERT [30] | 93.89±0.19 | 87.62±0.24 | 97.09±0.21 | 84.35±0.25 |
| | BLIP BERT [20] | 93.74±0.23 | 88.39±0.25 | 97.63±0.36 | 83.99±0.27 |
| | $\text{VBB}_{\text{u}}$ DeBERTa V3 | 98.16±0.03 | 88.11±0.07 | 98.28±0.04 | 91.31±0.11 |
| | $\text{VBB}_{\text{u}}$ DistilBERT | **99.76±0.02** | 99.63±0.03 | 99.54±0.03 | **99.64±0.03** |
| | $\text{VBB}_{\text{u}}$ BERT | 99.71±0.02 | **99.77±0.01** | **99.83±0.06** | 99.44±0.05 |
| | $\text{VBB}_{\text{u}}$ CLIP BERT | 99.73±0.01 | 99.72±0.01 | 99.51±0.00 | 99.47 ± 0.02 |

**Table 2.** Results for multimodal, unimodal and distilled (Dist) unimodal (Unim) models in VCTC on the Aqua datasets, measured using micro accuracy ($\text{Acc}_{\text{micro}}$), macro accuracy ($\text{Acc}_{\text{macro}}$), Area Under the ROC Curve (AUC) and the F1 score. For all models, performance is reported as the mean ± standard deviation over five independent runs. The best result for each metric is highlighted in bold and the second best result is underlined.

representations often match or exceed their multimodal counterparts, $\text{VBB}_{\text{u}}$ CLIP BERT achieves 84.46% macro accuracy and 91.24% AUC, outperforming $\text{VBB}_{\text{m}}$ CLIP BERT at 83.66% and 91.02% respectively. This suggests that our distillation process improving generalization on minority classes. On Aqua, $\text{VBB}_{\text{u}}$ results are closed to $\text{VBB}_{\text{m}}$ and VIKING, exceding 99% in all metrics. Architecture generalization is confirmed across backbones. Whether starting from BERT or DeBERTa, the VBB process yields improvements on Artpedia and on AQUA over text-only baselines. The persistent performance difference between datasets (Artpedia: 75-85% range, AQUA: 95-99% range) indicates that longer, more ambiguous art descriptions require richer representations than short, focused questions. In this context, the VBB process stands out both for its robustness on Artpedia, where it consistently ranks among the top performers, and for its scores on AQUA, where it achieves 99% results for all metrics, matching VIKING without requiring image processing at inference.

*Qualitative analysis.* For the 803 images common to Aqua and Artpedia, 7 contain documents incorrectly classified by the BERT model (see example in the table 3). VBB CLIP BERT was able to correctly label all these documents.

| Common images | Artpedia | AQUA |
|---|---|---|
|  | *"This scene is very similar to other paintings Ruisdael made of waterfalls."* Class = contextual Predicted Class = visual | *"what is an area which was a rich vein of inspiration for ruisdael?"* |
|  | *"A third version exists where the boy is grasping a whistle" or flute."* Class = contextual Predicted Class = visual | *"what is laughterl?"* |

**Table 3.** Examples of documents derived from common images in the AQUA and Artpedia datasets that were misclassified by BERT.

| | Artpedia | | | |
|---|---|---|---|---|
| | $Acc_{micro}$ | $Acc_{macro}$ | AUC | F1 score |
| BERT [9] | 79.46±0.30 | 83.32±0.15 | 90.57±0.08 | 75.32 ± 0.21 |
| $VBB_m$ CLIP BERT | **82.36**±**0.29** | 79.46±0.86 | 87.72±0.78 | 72.64 ± 1.00 |
| $VBB_u$ CLIP BERT | 81.84±0.34 | **84.33**±**0.40** | **90.96**±**0.37** | **76.95** ± **0.44** |

**Table 4.** Results for visual-contextual token classification on Artpedia dataset.

### 5.4 Token-level classification results

At token level, we aim to assess whether the model is capable of classifying each token of the text mention according to the document class. Within that context, Table 4 and Table 5 shows that $VBB_u$ CLIP BERT achieves the best token-level performance across both datasets (macro accuracy, AUC, F1), indicating that the adapted text encoder concentrates the probability mass on image-groundable tokens even without access to images at inference. Note that on Artpedia, BERT remains a strong second model, while $VBB_m$ CLIP BERT leads in micro accuracy (reflecting greater confidence on majority tokens) but trails on calibration-sensitive metrics consistent with the partial-groundability nature of VCTC.

To gain a deeper understanding of the models' behavior, we performed a granular analysis of token-level classification errors. Tables 24 present the most frequently misclassified words from both contextual and visual documents, on the Artpedia and AQUA datasets, for BERT and $VBB_u$ CLIP BERT. BERT misclassified a total of 4,829 unique tokens, with 4,169 tokens from contextual documents versus 660 from visual. This reflects the same class imbalance observed at the document level. We found that the most frequent errors for visual documents often involved names and concepts such as "magnificat" and "lamentation", which require external knowledge. Conversely, errors from contextual documents included tokens like "head", "flesh", and "brushstrokes", all of which carry strong perceptual connotations. $VBB_u$ CLIP BERT exhibits a behavior broadly similar, misclassifying 4,582 tokens (3,795 contextual and 787 visual), though the nature of its errors differs subtly. A closer inspection reveals that

| | AQUA | | | |
|---|---|---|---|---|
| | $Acc_{micro}$ | $Acc_{macro}$ | AUC | F1 score |
| BERT [9] | 84.08±0.50 | 89.18±0.35 | 91.74±0.32 | 76.41 ± 0.58 |
| $VBB_m$ CLIP BERT | 97.90±0.33 | 96.00±0.67 | 95.98±0.59 | 95.77 ± 0.69 |
| $VBB_u$ CLIP BERT | **99.72**±**0.01** | **99.72**±**0.00** | **99.13**±**0.00** | **99.47** ± **0.02** |

**Table 5.** Results for visual-contextual token classification on AQUA dataset.

| $VBB_u$ CLIP BERT | | BERT | |
|---|---|---|---|
| Ground truth label | | | |
| contextual | visual | contextual | visual |
| foreground | kalevipoeg | **mercury** | christmas |
| portrays | lamentation | **batarnay** | **beach** |
| goddess | botticelli | **bathsheba** | **tree** |
| **heterosexual** | 170x65 | **simplicity** | **couch** |
| **tóibín** | paston | **zacharias** | who |
| pearl | lennuk | **saskia** | sit |
| peasant | mcneill | devoid | next |
| persica | ghirlandaio | controted | to |
| concvlacaberis | metamorphoses | **solomon** | a |
| your | pareja | **sienese** | on |

**Table 6.** Top misclassified words for contextual and visual labels. Word in the columns were classified by the model as the opposite of the ground truth label. In bold words with high-confidence misclassification judged by human.

compared to BERT, it displays a more balanced prediction across tokens, suggesting a better handling of class ambiguity.

## 6 Conclusion and future work

In this work, we proposed a novel task, Visual-Contextual Text Classification (VCTC), to evaluate whether a text-based model can acquire a deeper visual understanding through multimodal pretraining. We found that our co-trained unimodal model, $VBB_u$, proved to be highly effective at this task. It outperformed text-only baselines and also consistently and effectively distinguished between visual and contextual tokens, even without access to the image during inference. Collectively, these findings underscore the potential of co-training strategies to enhance the visual knowledge of unimodal models. Our research provides a viable and effective direction for improving text-only models by leveraging multimodal pretraining, thereby eliminating the need for multimodal reasoning during the inference stage. While this work focuses on the art domain, a promising future direction is to evaluate the generalizability of our approach to other domains where text is implicitly grounded in visual information, such as product reviews, medical or real estate descriptions. This would test the robustness of visual knowledge conveyance in different contexts.

**Acknowledgements** The ABC project was managed by the Centre Hospitalier Universitaire de Caen Normandie, 14000 CAEN, France (Ringgold ID: 26962). It was funded by the GIP Millénaire Caen 2025 as part of the ABC project

(artbienetrecerveau.fr) and benefit from state aid managed by the National Research Agency under France 2030, CaeSAR (Caen, Stratégie pour l'Accélération en Recherche) (ANR-23-EXES-0001) and the Normandy region. Computing resources were provided by CRIANN.

## References

1. Aftan, S., Shah, H.: A survey on bert and its applications. In: 2023 20th Learning and Technology Conference (L&T). pp. 161–166. IEEE (2023)
2. Aladago, M.M.: Incorporating visual information into natural language processing (2025)
3. Alper, M., Fiman, M., Averbuch-Elor, H.: Is bert blind? exploring the effect of vision-and-language pretraining on visual language understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6778–6788 (2023)
4. Bai, Z., Nakashima, Y., Garcia, N.: Explain me the painting: Multi-topic knowledgeable art description generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5422–5432 (2021)
5. Becattini, F., Bongini, P., Bulla, L., Bimbo, A.D., Marinucci, L., Mongiovì, M., Presutti, V.: Viscounth: a large-scale multilingual visual question answering dataset for cultural heritage. ACM Transactions on Multimedia Computing, Communications and Applications **19**(6), 1–20 (2023)
6. Bongini, P., Becattini, F., Bagdanov, A.D., Del Bimbo, A.: Visual question answering for cultural heritage. In: IOP Conference Series: Materials Science and Engineering. vol. 949, p. 012074. IOP Publishing (2020)
7. Cetinic, E., She, J.: Understanding and creating art with ai: Review and outlook. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(2), 1–22 (2022)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
10. Fernando, K.R.M., Tsokos, C.P.: Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems **33**(7), 2940–2951 (2021)
11. Fields, J., Chovanec, K., Madiraju, P.: A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? IEEE Access **12**, 6518–6531 (2024)
12. Fu, J., Xu, S., Liu, H., Liu, Y., Xie, N., Wang, C.C., Liu, J., Sun, Y., Wang, B.: Cma-clip: Cross-modality attention clip for text-image classification. In: 2022 IEEE international conference on image processing (ICIP). pp. 2846–2850. IEEE (2022)
13. Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., Mitamura, T.: A dataset and baselines for visual question answering on art. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 92–108. Springer (2020)

14. Gasparetto, A., Marcuzzo, M., Zangari, A., Albarelli, A.: A survey on text classification algorithms: From text to predictions. Information **13**(2), 83 (2022)
15. He, P., Gao, J., Chen, W.: DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=sE7-XhLxHA
16. He, P., Liu, X., Gao, J., Chen, W.: Deberta: decoding-enhanced bert with disentangled attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=XPZIaotutsD
17. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International conference on machine learning. pp. 5583–5594. PMLR (2021)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
19. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
21. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
22. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=Bkg6RiCqY7
26. Ma, C., Shen, A., Yoshikawa, H., Iwakura, T., Beck, D., Baldwin, T.: On the (in) effectiveness of images for text classification. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 42–48 (2021)
27. Ma, C., Shen, A., Yoshikawa, H., Iwakura, T., Beck, D., Baldwin, T.: On the effectiveness of images in multi-modal text classification: an annotation study. ACM Transactions on Asian and Low-Resource Language Information Processing **22**(3), 1–19 (2023)
28. Peng, F., Yang, X., Xiao, L., Wang, Y., Xu, C.: Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. IEEE Transactions on Multimedia **26**, 3469–3480 (2023)

29. Qin, L., Wang, W., Chen, Q., Che, W.: Cliptext: A new paradigm for zero-shot text classification. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 1077–1088 (2023)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
31. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
32. Scius-Bertrand, A., Jungo, M., Vögtlin, L., Spat, J.M., Fischer, A.: Zero-shot prompting and few-shot fine-tuning: Revisiting document image classification using large language models. In: International Conference on Pattern Recognition. pp. 152–166. Springer (2024)
33. Sheng, S., Van Gool, L., Moens, M.F., Hinrichs, E.W., Hinrichs, M., Trippel, T.: A dataset for multimodal question answering in the cultural heritage domain. In: Proceedings of the COLING 2016 Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH). pp. 10–17. ACL (2016)
34. Stefanini, M., Cornia, M., Baraldi, L., Corsini, M., Cucchiara, R.: Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain. In: Image Analysis and Processing–ICIAP 2019: 20th International Conference, Trento, Italy, September 9–13, 2019, Proceedings, Part II 20. pp. 729–740. Springer (2019)
35. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. pp. 5099–5110. Association for Computational Linguistics (2019)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
37. Wang, P., Li, D., Hu, X., Wang, Y., Zhang, Y.: Clipmulti: Explore the performance of multimodal enhanced clip for zero-shot text classification. Computer Speech & Language **90**, 101748 (2025)
38. Wang, Z., Codella, N., Chen, Y.C., Zhou, L., Dai, X., Xiao, B., Yang, J., You, H., Chang, K.W., Chang, S.f., et al.: Multimodal adaptive distillation for leveraging unimodal encoders for vision-language tasks. arXiv preprint arXiv:2204.10496 (2022)
39. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
40. Zhang, C., Van Durme, B., Li, Z., Stengel-Eskin, E.: Visual commonsense in pretrained unimodal and multimodal models. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 5321–5335. Association for Computational Linguistics, Seattle, United States (2022)
41. Zhang, Y., Wang, P., Chen, Q., Zhou, J., Wang, Y., Li, M., Qin, L.: Labclip: Label-enhanced clip for improving zero-shot text classification. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11406–11410. IEEE (2024)

# Supplementary Materials

## A Experiments details

### A.1 Datasets

| Dataset | $\text{Documents}_{\text{v}}$ | $\text{Documents}_{\text{c}}$ | $\text{Tokens}_{\text{v}}$ | $\text{Tokens}_{\text{c}}$ |
|---|---|---|---|---|
| Artpedia [34] | 1,022 | 2,069 | 10,300 | 20,943 |
| AQUA [13] | 1,269 | 3,643 | 8,314 | 45,189 |

**Table 7.** Number of documents and tokens per label, visual ($_{\text{v}}$) or contextual ($_{\text{c}}$), in the test splits of Artpedia and AQUA.

*Artpedia.* The Artpedia dataset consists of 2,930 paintings, each with textual descriptions manually annotated as either visual (describing perceivable elements in the image) or contextual (requiring external knowledge). In total, the dataset comprises 9,173 visual and 19,039 contextual documents.

*AQUA.* The AQUA dataset includes 21,384 artworks where images paired with question-answer pairs, annotated to indicate whether the answer requires visual information from the image or external knowledge. The same image may appear multiple times with different questions. The full dataset contains 32,345 visual and 47,503 contextual questions.

### A.2 Extended Training/Implementation Details

*Global.* Table 8 consolidates the default optimization, batching, sequence, and aggregation choices shared across all experiments. These settings ensure fair comparisons (common optimizer, class-weighting, and pooling) and reproducibility, unless a model/dataset explicitly overrides them in later tables.

*Hyperparameters.* For Table 9, we list the small, log-scale grid used for validation selection (LR, WD, batch size, warmup, aggregator $\gamma$, loss weights), along with slots to record the chosen configuration per model/dataset. It provides how we pick hyperparameters (best validation loss, ...) before reporting test means over five seeds.

*Scheduling.* In Table 10 we specify, per model, which blocks are trainable at each stage and the corresponding LR/WD/epoch limits. $\text{VBB}_{\text{m}}$ uses a stabilizing two-stage regimen (frozen encoders $\rightarrow$ end-to-end), BERT follows standard end-to-end fine-tuning, and $\text{VBB}_{\text{u}}$ performs a brief head warmup before encoder fine-tuning to cleanly transfer multimodal priors into a text-only classifier.

| | **Default** | **Notes** |
|---|---|---|
| Optimizer | AdamW | $\beta_1$=0.9, $\beta_2$=0.999, |
| LR schedule | Warmup 10% | No decay unless stated |
| Early stopping | Patience 5 | Restore best val. loss |
| Grad clipping | 1.0 (global $\ell_2$) | Applied every step |
| Sequence length | $L_{\max} = 77$ | (BERT) masks for padding |
| Batch size | 16, 32 or 64 | See search grid |
| Token head | 1-layer MLP | Hidden size = encoder |
| Doc aggregator | Soft-majority | $\tau = \frac{1}{2}$, $\gamma = 10$ (default) |
| Hardware | 1×A100 80GB | CUDA 12.x, PyTorch 2.x |
| Seeds | {15, 42, 256, 484} | Mean±std over 5 runs |

**Table 8.** Default values unless noted.

| **Hyperparameter** | **Grid** |
|---|---|
| Learning rate | $\{1e^{-4}, 5e^{-4}, 1e^{-5}, 5e^{-5}, 1e^{-6}, 5e^{-6}, 1e^{-7}\}$ |
| Weight decay | $\{1e^{-4}, 1e^{-5}, 1e^{-6}, 1e^{-7}\}$ |
| Batch size | {16, 32, 64} |
| Warmup ratio | {0.1, 0.2} (abl.) |
| Aggregator $\gamma$ | {5, 10, 20} (abl.) |
| Loss weights $(\lambda_{\text{tok}}, \lambda_{\text{doc}})$ | $\{(1,1),(1,0),(0,1)\}$ |

**Table 9. Validation search grid and selection.** Best by val. loss (tie: macro-F1).

*$VBB_m$ training phase.* First stage of training $\text{VBB}_\text{m}$ consists to freeze encoders which initialize BERT and ViT from public checkpoints, freeze both backbones, and train the cross-attention fusion and token/document heads using weighted BCE with dynamic per-batch class weights. Token probabilities are $p_{j,k} = \sigma(\ell_{j,k})$ from an 1-layer MLP, and the document score uses soft-majority pooling (Equation 5 from Section 4.3). In the second step, we unfreeze all weights and continue optimizing the total loss $\mathcal{L}$ with early stopping. After co-training, we discard the image path and keep the adapted text encoder as $VBB_u$. Thus, a small classifier is attached, we warm up the head, then fine-tune the encoder under the same objectives but with text-only inputs. At inference, $VBB_u$ outputs token-level $p_{j,k}$ and a document decision $p_j = \sigma(\ell_j)$ via the same aggregator.

| Model | Trainable blocks | LR / WD / Max epochs |
|---|---|---|
| $VBB_m$ | Fusion+Heads (BERT/ViT frozen) | $1\times10^{-5}$ / $1\times10^{-5}$ / 2 |
| | All (BERT+ViT+Fusion+Heads) | $1\times10^{-5}$ / $1\times10^{-4}$ / 10 |
| BERT | All BERT layers + Head | $5\times10^{-5}$ / $1\times10^{-5}$ / 10 |
| $VBB_u$ | Classifier only | $5\times10^{-5}$ / $5\times10^{-5}$ / 4 |
| | Encoder + Head | $1\times10^{-5}$ / $1\times10^{-5}$ / 10 |

**Table 10.** Stage-wise LRs/WDs and freeze policies.

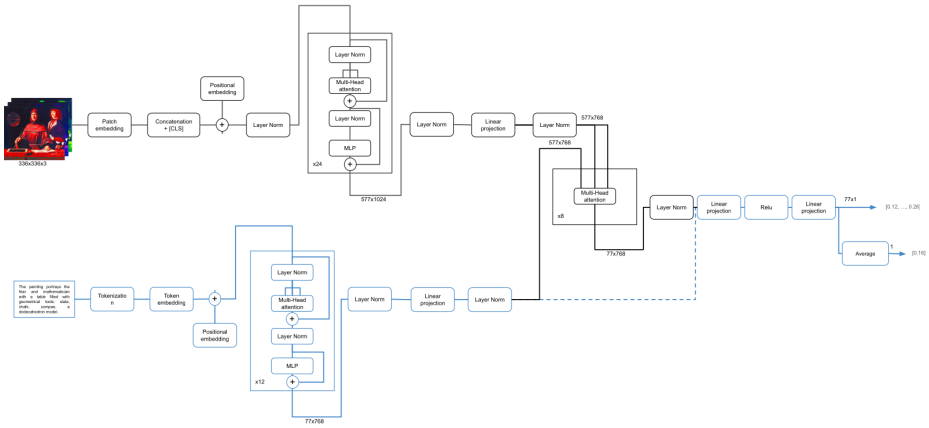


**Fig. 3.** Overview of the full ViT-Boosted BERT ($VBB_m$) multimodal architecture. The figure also presents, in blue, the unimodal $VBB_u$ architecture, which uses the text encoder from $VBB_m$.

**Algorithm 1** $VBB_m$ training (two-stage)

```
Initialize BERT, ViT (pretrained); fusion & heads (random)
for stage ∈ {frozen, end-to-end} do
    Freeze/unfreeze encoders accordingly; set lr, wd
    for epoch = 1, ..., E_max do
        for minibatch do
            Compute p_{j,k} via text/img cross-attention
            Aggregate to p_j (soft-majority)
            Compute L
            Backprop, clip grad ≤ 1.0, optimizer step
        end for
        Evaluate on validation and early stop if no improvement
    end for
end for
```

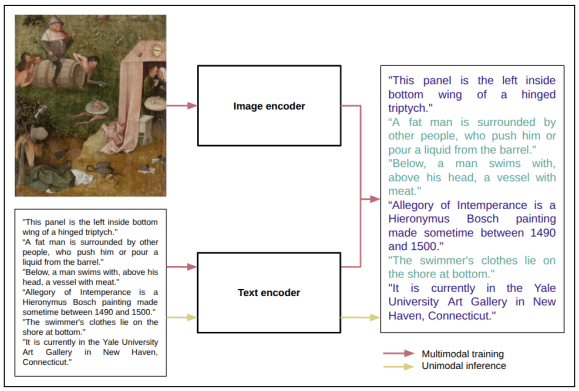


**Fig. 4.** A Language Model enriched with visual knowledge for VCTC in art description. The same text encoder is used for both multimodal training and unimodal inference. Green represents visual information and blue represents contextual information.

## B ViT-Boosted BERT Architecture

The complete multimodal architecture of $\text{VBB}_\text{m}$ model is shown in Figure 3. This model uses a Vision Encoder (ViT) and a Text Encoder (BERT) to process the input modalities. The visual and textual representations are then fused using a cross-attention mechanism, and the resulting representation is passed to a classification head. The unimodal $\text{VBB}_\text{u}$ model, which is highlighted in blue in the figure, is built on the text encoder from $\text{VBB}_\text{m}$, followed by a classification head.

Table 11 and table 12 expands on the results presented in the main paper by including additional model variants and recent architectures. For the multimodal comparison, we have added BLIP-2 and BLIP-3 (xGen-MM). In the unimodal section, we included various sizes of BERT base models (small, base, and large) like RoBERTa, providing a more comprehensive baseline comparison. This additional analysis reaffirms the effectiveness of our proposed method and highlights its competitive performance against a broader range of state-of-the-art models.

Table 14 present different result for LLaVA, following [32], under zero-shot and few-shot prompting. For zero-shot, we test three concise JSON-constrained prompts: PZ1, PZ2 and PZ3 (Table 13). We then select the best zero-shot prompt (PZ2) and progressively enrich it for few-shot. PD adds a brief visual/contextual description while PF1 augments PD with one simple example per class. Finally, PF2 further adds one hard example per class.

## C Additional details on ablation study

### C.1 Pre-training strategy

Table 15, unimodal models derived from efficient multimodal architectures ($\text{VBB}_\text{u}$ CLIP BERT dynamic) outperform both CLIP BERT and BLIP BERT, high-

| | | Artpedia | | | |
|---|---|---|---|---|---|
| | | $Acc_{micro}$ | $Acc_{macro}$ | AUC | F1 score |
| Unimodal | DeBERTa base [16] | 75.21 | 80.53 | 92.73 | 71.97 |
| | DeBERTa large [16] | 77.22 | 81.08 | 86.68 | 72.86 |
| | DeBERTa V3 small [15] | 65.32 | 71.84 | 83.74 | 63.46 |
| | DeBERTa V3 base [15] | 69.28 | 74.54 | 89.93 | 66.24 |
| | DeBERTa V3 large [15] | 68.42 | 74.52 | 88.51 | 66.16 |
| | DistilBERT [31] | 75.23 | 80.51 | **92.77** | 71.96 |
| | DistilRoBERTa [31] | 77.32 | 81.96 | 92.08 | 73.62 |
| | RoBERTa base [24] | 75.84 | 80.91 | 92.34 | 72.41 |
| | RoBERTa large [24] | 76.45 | 81.28 | 92.37 | 73.09 |
| | BERT [9] | **79.50** | **83.39** | 90.89 | **75.37** |
| Multimodal | CLIPText [29] | 82.91 | 84.20 | 91.18 | 77.73 |
| | CLIPText (Ensemble) [29] | 80.59 | 82.65 | 92.14 | 74.98 |
| | Prompt-CLIPText [29] | 83.28 | 83.81 | 90.05 | 77.19 |
| | Prompt-CLIPText (Ensemble) [29] | 83.88 | 84.49 | **92.34** | 78.00 |
| | LABCLIP [41] | 79.97 | 83.00 | 92.02 | 75.72 |
| | BLIP [20] | 74.72 | 78.82 | 88.74 | 70.39 |
| | BLIP 2 [18] | 76.35 | 81.09 | 90.62 | 72.67 |
| | xGen-MM (BLIP 3) [39] | 76.06 | 79.62 | 89.84 | 71.34 |
| | LLaVA [23] | 79.19 | 79.67 | 87.54 | 71.77 |
| | Prompt LLaVA [23] | 75.51 | 68.10 | _ | 55.14 |
| | ViLT [17] | 78.06 | 80.90 | 90.62 | 72.75 |
| | $VBB_m$ DeBERTa V3 | 78.96 | 77.72 | 85.88 | 69.97 |
| | $VBB_m$ DistilBERT | 80.50 | 82.69 | 90.79 | 75.23 |
| | $VBB_m$ DistilRoBERTa | 82.28 | 84.02 | 89.58 | 77.96 |
| | $VBB_m$ RoBERTa | 81.15 | 83.39 | 90.67 | 75.95 |
| | $VBB_m$ BERT | 83.61 | **84.93** | 91.31 | **78.18** |
| | $VBB_m$ CLIP BERT | **84.98** | 83.66 | 91.02 | 77.83 |
| Distilled | CLIP BERT [30] | 71.81 | 77.37 | 88.16 | 68.75 |
| | BLIP BERT [20] | 73.02 | 76.82 | 88.48 | 68.37 |
| | $VBB_u$ DeBERTa V3 | 76.59 | 78.63 | 86.95 | 70.51 |
| | $VBB_u$ DistilBERT | **84.10** | 84.41 | **91.65** | **78.05** |
| | $VBB_u$ DistilRoBERTa | 77.25 | 80.87 | 91.00 | 72.70 |
| | $VBB_u$ RoBERTa | 79.97 | 82.76 | 91.64 | 75.03 |
| | $VBB_u$ BERT | 83.25 | 84.16 | 91.49 | 77.46 |
| | $VBB_u$ CLIP BERT | 81.94 | **84.46** | 91.24 | 77.09 |

**Table 11.** Detailed results for the VCTC task on the Artpedia dataset. The best results for each metric are highlighted in bold, and the second-best results are underline.

lighting that multimodal pre-training allows the text encoder to capture more visual-contextual than standard visually-grounded text-only encoders.

### C.2 Loss weighting

As shown in Table 15, models trained with dynamic loss weighting consistently outperform those using static weighting. This demonstrates that dynamically adjusting the weight of classes during training is effective at addressing class imbalance and enhances the model's ability to distinguish between visual and contextual content. This explanation is valid for BLIP and VBB.

### C.3 Cross-attention layers

As illustrated in Figure 5 and detailed in Table 16, increasing the number of layers from 2 to 24 leads to an increase in accuracy and F1 score, which indicates that deeper cross-attention layers allow for a richer and more effective integration of visual and textual features. However, the AUC increases up to 8 layers and

| | | Aqua | | | |
|---|---|---|---|---|---|
| | | $Acc_{micro}$ | $Acc_{macro}$ | AUC | F1 score |
| Unimodal | DeBERTa base [16] | **97.01** | **90.54** | **98.65** | **86.87** |
| | DeBERTa large [16] | 94.33 | 88.72 | 97.56 | 85.89 |
| | DeBERTa V3 small [15] | 92.47 | 85.98 | 96.29 | 81.24 |
| | DeBERTa V3 base [15] | 94.89 | 87.68 | 97.02 | 86.41 |
| | DeBERTa V3 large [15] | 93.89 | 87.32 | 97.08 | 83.54 |
| | DistilBERT [31] | 95.35 | 90.48 | 98.32 | 86.39 |
| | DistilRoBERTa [31] | 95.24 | 89.37 | 98.31 | 86.25 |
| | RoBERTa base [24] | 95.33 | 89.59 | 98.32 | 86.33 |
| | RoBERTa large [24] | 95.07 | 89.49 | 98.00 | 86.28 |
| | BERT [9] | 84.90 | 89.76 | 92.14 | 77.37 |
| Multimodal | VIKING classifier [13] | 99.52 | 99.22 | **99.93** | 99.28 |
| | CLIPText [29] | 96.33 | 82.81 | 99.16 | 80.27 |
| | CLIPText (Ensemble) [29] | 95.76 | 82.89 | 99.51 | 79.25 |
| | Prompt-CLIPText [29] | 94.47 | 74.61 | 96.51 | 66.00 |
| | Prompt-CLIPText (Ensemble) [29] | 94.84 | 75.48 | 96.40 | 67.46 |
| | LABCLIP [41] | 96.54 | 83.78 | 96.49 | 80.42 |
| | BLIP [20] | 90.23 | 88.84 | 94.86 | 84.47 |
| | BLIP 2 [18] | 95.12 | 90.47 | 98.18 | 87.83 |
| | xGen-MM (BLIP 3) [39] | 94.71 | 89.23 | 97.91 | 86.13 |
| | ViLT [17] | 95.50 | 90.27 | 98.32 | 87.33 |
| | $VBB_m$ DeBERTa V3 | 98.88 | 88.07 | 98.29 | 91.86 |
| | $VBB_m$ DistilBERT | 95.62 | 91.54 | 98.54 | 90.75 |
| | $VBB_m$ BERT | 96.60 | 93.44 | 93.42 | 92.75 |
| | $VBB_m$ CLIP BERT | **99.88** | **99.85** | 99.52 | **99.77** |
| Dist | CLIP BERT [30] | 93.89 | 87.62 | 97.09 | 84.35 |
| | BLIP BERT [20] | 93.74 | 88.39 | 97.63 | 83.99 |
| | $VBB_u$ DeBERTa V3 | 98.16 | 88.11 | 98.28 | 91.31 |
| | $VBB_u$ DistilBERT | **99.76** | 99.63 | 99.54 | **99.64** |
| | $VBB_u$ BERT | 99.71 | **99.77** | **99.83** | 99.44 |
| | $VBB_u$ CLIP BERT | 99.73 | 99.72 | 99.51 | 99.47 |

**Table 12.** Detailed results for the VCTC task on the AQUA dataset. The best results for each metric are highlighted in bold, and the second-best results are underline.

then slightly decreases for the 12 and 24 layers. While models with more layers achieve the best F1 and accuracy scores, this slight drop in AUC suggests a compromise between overall classification performance and the model's ability to effectively distinguish between classes.

### C.4 Vision embedding

We isolate the contribution of the visual branch by swapping the image encoder (ResNet-50 vs. ViT-L/14) while fixing the CLIP-BERT text backbone. Table 17 reports both the multimodal stage ($VBB_m$ CLIP BERT) and the final text-only model ($VBB_u$ CLIP BERT). ViT-L/14 delivers a consistent AUC advantage, around +4.7 in both stages, and a small gain in $Acc_{micro}$ for the multimodal model. After removing the image modality, differences narrow: ResNet-50 is marginally higher on $Acc_{macro}$ and F1, while ViT-L/14 retains a clear AUC lead. Overall, the vision encoder choice is not the dominant factor for the text-only classifier, and both variants perform similarly on accuracy/F1. Though ViT-L/14 is preferable when ranking quality (AUC) matters.

| Prompt 1 (PZ1) | Prompt 2 (PZ2) | Prompt 3 (PZ3) |
|---|---|---|
| `You are a text classifier. You must classify a text into one of the following categories: "visual" or "contextual". Your answer must follow the schema below, including the original sentence and the category you chose. JSON schema: {"possible_categories": ["visual", "contextual"], "output": { "sentence": "original_sentence", "categorie_classifie": "one_of_the_possible_ categories" } } Sentence to classify: "{sentence}" JSON answer:` | `Classify a text in one of these category: "visual" or "contextual". To answer, follow the schema below. You must include the original sentence and the chosen category. JSON schema: { "possible_categories": ["visual", "contextual"], "output": { "sentence": "original_sentence", "categorie_classifie": "one_of_the_possible_ categories" } } Sentence to classify: "{sentence}" JSON answer:` | `Your task is to classify a text. Select one appropriate category for the text between: "visual" or "contextual". Follow the schema below to answer. JSON schema: { "possible_categories": ["visual", "contextual"], "output": { "sentence": "original_sentence", "categorie_classifie": "one_of_the_possible_ categories" } } Sentence to classify: "{sentence}" JSON answer:` |

**Table 13.** Presentation of the three prompts used for zero-shot prompting in LLaVA

| | Artpedia | | | |
|---|---|---|---|---|
| | $\text{Acc}_{\text{micro}}$ | $\text{Acc}_{\text{macro}}$ | F1 score | Invalid % |
| LLaVA zero-shot (PZ1) | 66.20± 0.72 | 65.37± 0.67 | 56.63± 0.77 | 11.41 ± 0.27 |
| LLaVA zero-shot (PZ2) | 68.64 ± 0.30 | 68.34± 0.45 | 58.50 ± 0.31 | 6.18 ± 0.36 |
| LLaVA zero-shot (PZ3) | 64.44 ± 0.42 | 63.50 ± 0.62 | 54.31 ± 0.85 | 7.16 ± 0.29 |
| LLaVA descriptive (PD) | 70.63± 0.21 | 59.45± 0.35 | 34.47± 0.89 | 7.57± 0.10 |
| LLaVA 1-shot (PF1) | 75.51± 0.47 | 68.10± 0.91 | 55.14± 1.76 | 25.34 ± 0.71 |
| LLaVA 2-shot (PF2) | 73.99± 0.86 | 65.25 ± 0.89 | 49.11 ± 0.98 | 29.81 ± 0.27 |
| LLaVA classification | 79.19 ±0.12 | 79.67 ±0.05 | 87.54 ±0.11 | 71.77 ±0.07 |
| $\text{VBB}_{\text{u}}$ DistilBERT | **84.10 ± 0.14** | **84.41** ± 0.67 | **78.05 ± 0.12** | - |

**Table 14.** LLaVA details result on prompting

### C.5 Inference efficiency

Table 18 confirms the core promise of unimodal part: we use images only to teach the text encoder, then keep a text-only model with the same footprint as its original backbone. As expected, unimodals models are heavier than text baselines because of the vision branch ($\text{VBB}_{\text{m}}$ DistilBERT), while still faster than BLIP (28.78 ms) and leaner than LLaVA (318.37 ms). They are costlier than a compact VLM like LABCLIP. Dropping the image path at inference slashes cost: DistilBERT runs in 3.28 ms with 67 M parameters and 271 Inf w, almost 5.8 time faster with 5.5 time fewer params, and being 5.4 time lighter than its own multimodal counterpart. It is also 3.7 time faster and 2.2 time lighter than LABCLIP, and 97 time faster and 36 time lighter than LLaVA. Other unimodal variants retain similarly small footprints (BERT/CLIP-BERT/BLIP), with DeBERTa-V3 remaining the heaviest among our observed models. Finally, unimodal delivers order-of-magnitude wall-clock and memory savings over large VLMs, while preserving the accuracy gains learned in multimodal mode, making it practical for large-scale systems. Moreover, since training occurs primarily in

| | | Artpedia | | | |
|---|---|---|---|---|---|
| | | $Acc_{micro}$ | $Acc_{macro}$ | AUC | F1 score |
| Multimodal | $VBB_m$ CLIP BERT dynamic | **84.98** | **83.66** | **91.02** | **77.83** |
| | BLIP dynamic | 75.77 | 79.26 | 89.45 | 70.23 |
| | $VBB_m$ CLIP BERT static | 81.30 | 81.90 | 84.96 | 74.74 |
| | BLIP static | 75.74 | 79.53 | 89.67 | 70.48 |
| | $VBB_m$ CLIP BERT ResNet | 84.30 | 83.03 | 86.28 | 76.96 |
| Distilled | $VBB_u$ CLIP BERT dynamic | **81.94** | 84.46 | **91.24** | 77.09 |
| | BLIP BERT dynamic | 73.76 | 78.16 | 87.71 | 68.92 |
| | $VBB_u$ CLIP BERT static | 75.77 | 80.56 | 83.74 | 72.10 |
| | BLIP BERT static | 72.89 | 77.88 | 86.94 | 68.58 |
| | $VBB_u$ CLIP BERT ResNet | 81.78 | **84.69** | 86.53 | **77.20** |
| | CLIP BERT | 71.81 | 77.37 | 88.16 | 68.75 |
| | BLIP BERT | 73.02 | 76.82 | 88.48 | 68.37 |

**Table 15.** Influence of the different model components on document classification for the Artpedia dataset.

| | Artpedia | | | |
|---|---|---|---|---|
| | $Acc_{micro}$ | $Acc_{macro}$ | Auc | F1 score |
| $VBB_m$ CLIP BERT (2 layers) | 81.69 | 83.82 | 88.25 | 76.49 |
| $VBB_m$ CLIP BERT (4 layers) | 82.30 | 83.96 | 89.34 | 76.85 |
| $VBB_m$ CLIP BERT (8 layers) | 84.98 | 83.66 | **91.02** | 77.83 |
| $VBB_m$ CLIP BERT (12 layers) | 85.09 | 84.45 | 90.54 | 78.55 |
| $VBB_m$ CLIP BERT (24 layers) | **85.73** | **85.43** | 89.43 | **79.67** |

**Table 16.** $VBB_m$ CLIP BERT performance on the Artpedia dataset with varying numbers of cross-attention layers.

multimodal mode with only a lightweight classification layer to train in the unimodal phase, training time remains reasonable.

### C.6 Approach's components for token classification

Table 19 presents a detailed ablation study focused on the token classification task. The observations on the influence of model components are similar with the main paper analysis of document classification.

We found that the choice of image encoder has a negligible influence on final performance. While the main paper highlighted a marginal improvement with the ViT encoder for document classification, we observe a negligible performance gain when using the R-CNN encoder for the token classification task.

### C.7 Document-level error analysis

To gain a deeper understanding of our model's behavior, we conducted an analysis of misclassified documents. Table 20 reveal that BERT struggles more with contextual descriptions than with visual ones across both the Artpedia and AQUA datasets. This is likely due to the less ambiguous nature of purely visual documents. For instance, in "Little Blue Horse", the model misclassifies as contextual document. The sentence, while providing external knowledge, contains

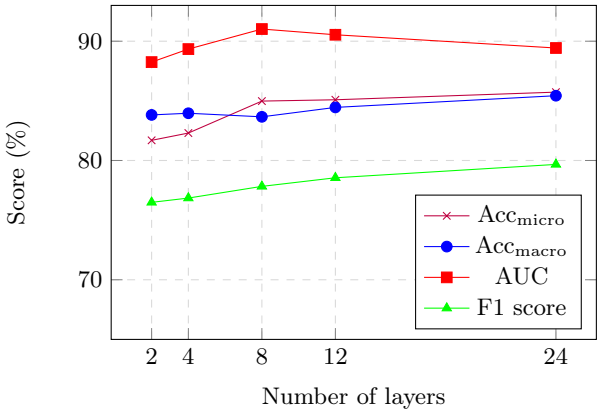


**Fig. 5.** Impact of the number of cross-attention layers on model performance, wrt. accuracy, AUC and F1 score.

| | Artpedia | | | |
|---|---|---|---|---|
| | Acc$_{micro}$ | Acc$_{macro}$ | AUC | F1 score |
| CLIP ViT | **84.98** | 83.66 | 91.02 | **77.83** |
| CLIP ResNet | 84.30 | 83.03 | 86.28 | 76.96 |
| CLIP BERT ViT | 81.94 | 84.46 | **91.24** | 77.09 |
| CLIP BERT ResNet | 81.78 | **84.69** | 86.53 | 77.20 |

**Table 17.** Influence of vision encoder

terms that can evoke a visual concept, confusing the model. Similarly, words such as "painting" or "canvas" refer words present in both visual documents and contextual ones, presenting a challenge for a text-only model.

The VBB$_u$ CLIP BERT model exhibits similar misclassification patterns on ambiguous documents (see Table 21). However, it demonstrates greater accuracy on certain cases that BERT fails on, such as "The work is stylistically..." and "The painting is an oil...". It also shares a common failure point with BERT on the sentence "Who is doctrinal message?", suggesting a similar limitation in handling certain types of highly abstract questions. Overall, VBB$_u$ CLIP BERT proves to be more robust on contextual descriptions, misclassifying fewer such documents. While it maintains only a moderate drop in performance on visual documents (17 documents, or 0.55% worse on Artpedia; 2 documents, or 0.04% worse on AQUA), its superior performance on contextual descriptions indicates a stronger grasp of the visual-contextual distinction. This comes at a small cost to visual classification precision.

### C.8 Analysis of misclassification on shared images

Table 22 highlights two cases where documents associated with the same image in both the Artpedia and AQUA datasets were misclassified by BERT alone. In fact, among the 803 images shared between the two datasets, BERT was the only

| | | Inf t | Param | Inf w |
|---|---|---|---|---|
| Unimodal | DeBERTa V3 [15] | 12.86 | 184.42 | 728.09 |
| | DistilBERT [31] | **3.28** | **66.96** | **270.93** |
| | DistilRoBERTa [31] | 3.56 | 82.12 | 328.94 |
| | RoBERTa [24] | 6.05 | 124.65 | 492.52 |
| | BERT [9] | 5.91 | 109.48 | 434.52 |
| Multimodal | LABCLIP [41] | **12.09** | **151.58** | **591.56** |
| | BLIP [20] | 28.78 | 223.35 | 2318.61 |
| | LLaVA [23] | 318.37 | 354.32 | 9792.34 |
| | $VBB_m$ DeBERTa V3 | 28.72 | 487.60 | 1903.84 |
| | $VBB_m$ DistilBERT | 19.13 | 370.13 | 1454.98 |
| | $VBB_m$ DistilRoBERTa | 19.30 | 385.30 | 1512.99 |
| | $VBB_m$ RoBERTa | 21.79 | 427.82 | 1675.22 |
| | $VBB_m$ BERT | 21.77 | 412.66 | 1617.22 |
| | $VBB_m$ CLIP BERT | 21.77 | 412.66 | 1617.22 |
| Dist | CLIP BERT [30] | 5.91 | 109.48 | 434.52 |
| | BLIP BERT [20] | 5.91 | 109.48 | 434.52 |
| | $VBB_u$ DeBERTa V3 | 12.86 | 184.42 | 728.09 |
| | $VBB_u$ DistilBERT | **3.28** | **66.96** | **270.93** |
| | $VBB_u$ CLIP BERT | 5.91 | 109.48 | 434.52 |

**Table 18.** Efficiency comparison of baselines against multimodal and unimodal models.

| | | Artpedia | | | |
|---|---|---|---|---|---|
| | | $Acc_{micro}$ | $Acc_{macro}$ | Auc | F1 score |
| Multimodal | $VBB_m$ CLIP BERT dynamic | **82.36** | **79.46** | **87.72** | **72.64** |
| | BLIP dynamic | 75.77 | 78.26 | 84.19 | 70.03 |
| | $VBB_m$ CLIP BERT static | 80.40 | 79.27 | 86.37 | 71.93 |
| | BLIP static | 75.67 | 78.78 | 83.67 | 70.29 |
| | $VBB_m$ CLIP BERT R-CNN | 81.35 | 77.63 | 81.96 | 70.27 |
| Distilled | $VBB_u$ CLIP BERT dynamic | 81.84 | 84.33 | **90.96** | 76.95 |
| | BLIP BERT dynamic | 73.76 | 77.16 | 82.81 | 68.72 |
| | $VBB_u$ CLIP BERT static | 75.94 | 80.53 | 85.21 | 72.11 |
| | BLIP BERT static | 72.89 | 76.88 | 80.81 | 68.38 |
| | $VBB_u$ CLIP BERT R-CNN | **81.88** | **84.55** | 89.94 | **77.13** |
| | CLIP BERT | 71.81 | 77.37 | 82.48 | 68.75 |
| | BLIP BERT | 73.02 | 76.82 | 81.73 | 68.34 |

**Table 19.** Influence of model components on token classification in Artpedia

model to produce classification errors. This occured for 7 images for a total of 44 misclassified documents (34 from Artpedia, 10 from AQUA)[3]. All these errors correspond to sentences that are primarily contextual, yet BERT predicted them as visual. In contrast, $VBB_u$ CLIP BERT and $VBB_m$ CLIP BERT were able to correctly label these images, suggesting successful visual information conveyance into the text-encoder.

### C.9 Token-level error analysis

To gain a deeper understanding of the models' behavior, we performed a granular analysis of token-level classification errors. Tables 23 and 24 present the most frequently misclassified words from both contextual and visual documents, on the Artpedia and AQUA datasets, for BERT and $VBB_u$ CLIP BERT.

BERT misclassified a total of 4,829 unique tokens. The misclassification of 4,169 tokens from contextual documents versus only 660 from visual ones reflects the same class imbalance observed at the document level. We found that

[3] There may exist different text descriptions for a given image.

| Artpedia | | AQUA | |
|---|---|---|---|
| from contextual | from visual | from contextual | from visual |
| – *"Little Blue Horse is an abstract painting by Franz Marc, from 1912."*<br>– *"The painting is an oil on canvas, with dimensions 57.5 x 73 centimeters."*<br>– *"The blue and ultramarine paint may be later additions."* | – *"The triptych is a portrait of the moments before Dyer's death from an overdose of pills in their hotel room."*<br>– *"The work is stylistically more static and monumental than Bacon's earlier triptychs of Greek figures and friends' heads."* | – *when did honthorst and terbrugghen spend around 1621?"*<br>– *"who did vincenzo giustiniani purchase?"*<br>– *"who is a children's game played as early as 2,000 years ago in greece?"*<br>– *"what made zurbarn famous?"* | – *"what stand next to a building?"*<br>– *"what do a group of people sit on a beach next to?"* |

**Table 20.** Documents misclassified by BERT for the Aqua and Artpedia datasets. *From contextual* refers to contextual sentences that were classified as visual by the model, while *from visual* refers to the inverse.

the most frequent errors for visual documents often involved names and concepts such as "magnificat" and "lamentation", which require external knowledge. Conversely, errors from contextual documents included tokens like "head", "flesh", and "brushstrokes", all of which carry strong perceptual connotations.

$VBB_u$ CLIP BERT exhibits a behavior broadly similar to BERT, misclassifying 4,582 tokens (3,795 contextual and 787 visual), though the nature of its errors differs subtly. A closer inspection reveals that the model mislabeled tokens as visual in 14 contextual documents and as contextual in 78 visual documents. Compared to BERT, it displays a more balanced prediction across tokens, suggesting a better handling of class ambiguity.

On AQUA, both models made fewer errors. BERT misclassified 2,335 contextual tokens as visual while incorrectly labeling only 13 visual tokens as contextual. The most frequent errors included proper names like "mercury" and "solomon" being mistakenly treated as visual entities. For $VBB_u$ CLIP BERT , the error counts were even lower, with only 71 contextual tokens and 19 visual tokens misclassified. Unlike Artpedia where the majority of misclassified words are ambiguous, only a portion of $VBB_u$ CLIP BERT's errors can be considered semantically valid. For instance, in one instance of fully mislabeled tokens, words like "tobacco", "bottle", and "hand" were classified as visual. While the wanted label was contextual, the tokens themselves describe perceptual features, indicating the model is making fine-grained distinctions that align with a visual understanding.

### C.10 Analysis of document-level classification

In Table 25, we aim at understanding the correlation between models in their classification capacities. Resuts highlight notable differences between the two datasets. On AQUA, only 2 documents (0.04%) were misclassified by all models,

| Artpedia | | AQUA | |
|---|---|---|---|
| from contextual | from visual | from contextual | from visual |
| – *"The pavement is decorated with marble tarsia, already used by Fra Angelico in earlier works such as the San Pietro Martire Triptych (1428-1429)"*<br>– *"The attempt resulted in a disaster, with the larger part of the Greeks slain."* | – *"The background was painted in gold, with a blending effect in correspondence of the saints' aureolas."*<br>– *"The details of the ruins are precisely observed as they appear at the site in Rome."*<br>– *"A storm is coming on."*<br>– *"The studio is ascetically bare."* | – *"what do the four eldest family members respond with invocatory verses from the book of psalms on?"*<br>– *"who is doctrinal message?"*<br>– *"what do such works as the young woman composing music present a rarefied view of?"* | – *"what walk down a dirt road?"*<br>– *"what is the table made of?"*<br>– *"where do a group of people sit?"*<br>– *"what stand next to each other?"* |

**Table 21.** Documents misclassified by $VBB_u$ CLIP BERT for the Aqua and Artpedia datasets. *From contextual* refers to contextual sentences that were classified as visual by the model, while *from visual* refers to the inverse.

whereas on Artpedia this number rises up to 447 documents (14.46%). This discrepancy illustrates that Artpedia is a more challenging corpus, likely due to longer and more contextually complex descriptions, compared to the shorter, more structured questions of AQUA.

Examining documents correctly classified by exactly two models reveals further patterns of model convergence. On Artpedia, BERT and $VBB_u$ CLIP BERT demonstrate better convergence, with 137 (4.43%) shared documents correctly classified. This compares to only 33 (1.07%) shared documents between the BERT and $VBB_m$ CLIP BERT, and 76 (2.46%) between $VBB_m$ CLIP BERT and $VBB_u$ CLIP BERT. On AQUA, the performance aligns with theoretical expectations derived from the relationship between $VBB_m$ CLIP BERT and $VBB_u$ CLIP BERT. Notably, 732 (14.90%) questions were accurately classified by both $VBB_m$ CLIP BERT and $VBB_u$ CLIP BERT, compared to only 3 (0.06%) between either $VBB_m$ CLIP BERT and BERT, or $VBB_u$ CLIP BERT and BERT. This divergence suggests that understanding complex documents requires a deeper analysis of syntactic relationships, a capability intrinsic to language models, while shorter questions can be effectively processed by multimodal architectures that are less dependent on deep text comprehension.

Furthermore, on Artpedia, 196 (6.31%) documents were correctly classified by only one model, with 40 (1.29%) by BERT, 65 (2.10%) by $VBB_u$ CLIP BERT, and 91 (2.94%) by $VBB_m$ CLIP BERT. These results indicate that $VBB_u$ CLIP BERT developed classification mechanisms distinct from its parent model. A similar trend was observed for the AQUA dataset, where each unimodal model correctly classified 1 (0.02%) question that others misinterpreted, versus 9 (0.18%) for $VBB_m$ CLIP BERT.

| Common images | Artpedia | AQUA |
|---|---|---|
|  | *"This scene is very similar to other paintings Ruisdael made of waterfalls."*<br>Class = contextual<br>Predicted Class = visual | *"what is an area which was a rich vein of inspiration for ruisdael?"* |
|  | *"A third version exists where the boy is grasping a whistle" or flute."*<br>Class = contextual<br>Predicted Class = visual | *"what is laughterl?"* |

**Table 22.** Examples of documents derived from common images in the AQUA and Artpedia datasets that were misclassified by BERT.

| Artpedia | | AQUA | |
|---|---|---|---|
| Ground truth label | | | |
| contextual | visual | contextual | visual |
| head | magnificat | **mercury** | christmas |
| foreground | lamentation | **batarnay** | **beach** |
| flesh | 170x65 | **bathsheba** | **tree** |
| depicted | palavas | **simplicity** | **couch** |
| portrays | lennuk | **zacharias** | who |
| brushstrokes | mcneill | **saskia** | sit |
| goddess | reminiscent | devoid | next |
| **heterosexual** | **candle** | contorted | to |
| **tóibín** | metamorphoses | **solomon** | a |
| seated | pareja | **sienese** | on |

**Table 23.** Top BERT misclassified words for contextual and visual labels. Word in the columns were classified by BERT as the opposite of the ground truth label. In bold words with high-confidence misclassification judged by human.

| Artpedia | | AQUA | |
|---|---|---|---|
| Ground truth label | | | |
| contextual | visual | contextual | visual |
| foreground | kalevipoeg | strewn | **skateboard** |
| portrays | lamentation | leaning | posing |
| goddess | botticelli | tavern | **forest** |
| **heterosexual** | 170x65 | run | for |
| **tóibín** | paston | parallel | **mirror** |
| pearl | lennuk | rear | where |
| peasant | mcneill | sitting | front |
| persica | ghirlandaio | window | picture |
| concvlacaberis | metamorphoses | **louvre** | who |
| your | pareja | people | in |

**Table 24.** Top $VBB_u$ CLIP BERT misclassified words for contextual and visual labels. Word in the columns were classified by $VBB_u$ CLIP BERT as the opposite of the ground truth label. In bold words with high-confidence misclassification judged by human.

| Correctly classified documents | AQUA | Artpedia |
|---|---|---|
| BERT ∩ $VBB_u$ CLIP BERT ∩ $VBB_m$ CLIP BERT | 4161 | 2202 |
| BERT ∩ $VBB_u$ CLIP BERT | 3 | 137 |
| BERT ∩ $VBB_m$ CLIP BERT | 3 | 33 |
| $VBB_u$ CLIP BERT ∩ $VBB_m$ CLIP BERT | 732 | 76 |
| BERT | 1 | 40 |
| $VBB_u$ CLIP BERT | 1 | 65 |
| $VBB_m$ CLIP BERT | 9 | 91 |
| Misclassified documents by BERT ∧ $VBB_u$ CLIP BERT ∧ $VBB_m$ CLIP BERT | 2 | 447 |
| Total | 4912 | 3091 |

**Table 25.** Documents correctly classified by one, two, or all three models, including BERT, $VBB_u$ CLIP BERT, and $VBB_m$ CLIP BERT.