

GLaDViT: interpretable gated fusion of frozen vision, text and layout features for document image classification

Murat Simsek

`murat.simsek@ostimteknik.edu.tr`

Ostim Technical University

Research Article

Keywords: Document image classification, Multimodal fusion, Gated fusion, Layout analysis, Interpretability, Tobacco-3482

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11167180/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

GLaDViT: interpretable gated fusion of frozen vision, text and layout features for document image classification

Murat Şimşek^{1*}

^{1*}Department of Artificial Intelligence Engineering, Faculty of Engineering, Ostim Technical University, Ankara, Türkiye.

Corresponding author(s). E-mail(s): murat.simsek@ostimteknik.edu.tr;

Abstract

Classifying scanned pages by document type is a routing step in enterprise document processing. Layout-aware models reach high accuracy but rely on large-scale document pretraining and full fine-tuning, and seldom expose why a decision was made. We study a deliberately lightweight alternative, GLaDViT, which keeps CLIP and BERT encoders frozen, derives a layout descriptor from Tesseract hOCR word boxes, and trains only a gated fusion head. Because the classifier is linear in the gated sum of modality embeddings, every class logit decomposes exactly into per-modality terms, which yields an attribution that is faithful to the computation by construction, unlike the gate weights alone. On a label-corrected release of Tobacco-3482 (2737 pages, fixed 2186/272/279 split), GLaDViT reaches $93.7 \pm 0.4\%$ accuracy over three seeds (95% Wilson interval 90.2–96.0%), against 87.7% for the best single modality; adding layout to vision and text contributes 2.8 accuracy points and gating a further 1.1. The layout stream nevertheless receives at most about 10% of the gate mass, indicating that gate weights are mixing coefficients rather than importance scores. A triple InfoNCE alignment of the three streams, used as a warm-up on frozen features, lowers accuracy to 90.3%; we analyse this negative result. Code and hOCR tooling will be made publicly available.

Keywords: Document image classification, Multimodal fusion, Gated fusion, Layout analysis, Interpretability, Tobacco-3482

1 Introduction

Enterprises and public bodies digitise large archives of scanned documents, and a recurring first step is to recognise the *type* of each page—letter, memo, form, invoice, report, and so on—so that it can be routed, indexed and processed. On the standard benchmarks RVL-CDIP [1] and Tobacco-3482 [2], the strongest results come from layout-aware transformers pretrained on millions of pages [3–5] or, for the small Tobacco-3482 set, from networks first trained on RVL-CDIP [6, 7]. These models are accurate but costly to adapt,

and they rarely report *why* a page was assigned to a class, which limits their use where decisions must be audited.

A cheaper operating point is to use general-purpose foundation encoders as frozen feature extractors and to train only a small head. Frozen encoders are inexpensive to train on, easy to update, and keep the pipeline auditable because the OCR front end [8] remains explicit. Three questions then arise. (a) How much does explicit OCR geometry add to frozen image [9] and text [10] embeddings, which already encode coarse layout? (b) Can the fusion be interpreted with

an attribution that is faithful to the computation rather than approximated post hoc [11, 12]? (c) Does multimodal contrastive alignment, which is effective when encoders are pretrained [9, 13, 14], help when the encoders cannot adapt?

We address these questions with GLaDViT (*Gated, Layout-aware Document classifier on frozen ViT and BERT features*). Our contributions are: (i) a parameter-efficient three-stream design—frozen CLIP ViT-B/32 and BERT-base embeddings plus a 262-dimensional layout descriptor computed from hOCR word boxes—combined by a normalised gating module, with an ablation that isolates the contributions of layout and gating and reports interval estimates; (ii) an exact, logit-level modality attribution for gated fusion with a linear head, together with evidence that raw gate weights, which are often reported as explanations, understate the contribution of the layout stream; (iii) an analysis of triple InfoNCE alignment under frozen backbones, reported as a negative result; and (iv) a sourced comparison with published Tobacco-3482 results that separates label versions and in-domain (RVL-CDIP) transfer. We do not claim state-of-the-art accuracy; the aim is to characterise a cheap and transparent operating point.

2 Related work

Document image classification. RVL-CDIP (400k pages, 16 classes) and Tobacco-3482 (3482 pages, 10 classes) are both drawn from the IIT-CDIP collection [1, 2, 15]. Tobacco-3482 results are usually reported under the protocol of Kang et al. [16]: 100 pages per class for training and validation, the rest for testing, over ten random partitions. Under this protocol, CNNs initialised on ImageNet reach 75–78%, whereas initialisation on RVL-CDIP—an in-domain source—raises accuracy above 90% [6]. Without RVL-CDIP transfer, the strongest image-based results we found are 90.1% for DocXClassifier [7] and 91.1% when OCR text is rendered into the image (Word-VIS) [17]. Multimodal networks fuse CNN and text features by feature concatenation or score ensembling [18, 19], and VLCDoC learns cross-modal document representations with intra- and inter-modality contrastive objectives [14]. Label quality has recently come under scrutiny: Lim et al. report that 11.7% of Tobacco-3482 pages are

mislabelled or fall outside the ontology and 16.7% admit several valid labels [20], and a comparable audit of RVL-CDIP identifies label errors and other evaluation pitfalls [21]. The label version and protocol must therefore be stated when numbers are compared.

Layout-aware multimodal models. LayoutLM and its successors [3–5], DocFormer [22] and LiLT [23] jointly pretrain text, image and 2-D position; UDOP [24] unifies the three modalities in a generative framework. The six-dimensional box encoding $(x_0, y_0, x_1, y_1, w, h)$ that we use follows this line of work [3]. These models rely on large-scale pretraining and full fine-tuning; GLaDViT instead trains only a fusion head on frozen features.

OCR-free and LLM-based understanding. Donut [25], Pix2Struct [26] and mPLUG-DocOwl [27] remove the OCR stage, and recent multimodal large language models extend this line to text-rich and multi-page documents. These directions trade an explicit, auditable OCR stage for scale; we retain the OCR front end.

Multimodal fusion and multi-way contrastive learning. Fusion strategies range from early concatenation and late score averaging to tensor fusion [28] and gated fusion [29]; see [30] for a survey. Gated multimodal units (GMU) learn input-dependent gates that weight modality-specific representations [29]; the gate of GLaDViT is a normalised, self-conditioned variant of this idea (Section 3.4). For representation alignment, CLIP optimises a symmetric InfoNCE loss [31] between two modalities [9]; ImageBind binds six modalities through image-paired InfoNCE losses [13] and TupleInfoNCE contrasts tuples of modalities [32]. Our triple objective, a weighted sum of the three pairwise losses, is an instance of this family; our interest is whether it helps when the encoders are frozen.

Interpretability. Post-hoc methods such as LIME [11] and SHAP [12] approximate a trained model after the fact. Whether attention or gate weights constitute explanations is debated [33, 34], and faithfulness—how accurately an explanation reflects the model’s actual reasoning—has to be demonstrated rather than assumed [35]. MM-SHAP quantifies modality contributions with Shapley values [36], and DocXClassifier produces attention-like maps for document images [7]. GLaDViT provides an exact additive modality

decomposition at no extra cost, which we contrast with the gate weights.

3 Materials and methods

3.1 Problem formulation

A document is a pair $D = (I, \mathcal{W})$ with page image I and OCR words $\mathcal{W} = \{(w_i, b_i)\}_{i=1}^N$, where w_i is a word and $b_i \in \mathbb{R}^4$ its bounding box. GLaD-ViT computes a fused representation $z \in \mathbb{R}^d$ from which a linear head predicts the class distribution $p(y \mid D)$ over C types. Fig. 1 shows the architecture.

3.2 Dataset

Tobacco-3482 [2] contains 3482 pages in ten classes (advertisement, email, form, letter, memo, news, note, report, resume, scientific). We use a publicly available label-corrected release¹ that applies the re-annotation of Lim et al. [20]: labels are replaced by the corrected ones, pages with multiple valid labels or without a valid label are removed, and page images are distributed with OCR text and Tesseract hOCR word boxes. The release contains 2737 pages with a fixed train/validation/test split of 2186/272/279; the 745 removed pages are therefore the ambiguous or out-of-ontology pages identified in [20]. Because label cleaning removes ambiguous pages, and because about a third of a strong model’s errors on the original labels have been traced to such pages [20], accuracies on this release are not directly comparable with results on the original labels. No RVL-CDIP data are used at any stage, so the model receives no in-domain transfer from IIT-CDIP.

3.3 Modality encoders

Vision. The page image is encoded by a frozen CLIP ViT-B/32 backbone [9, 37] at its native input resolution of 224×224 pixels, giving the 512-dimensional image embedding h_v .

Text. The OCR text is encoded by frozen BERT-base [10]; inputs longer than BERT’s limit of 512 WordPiece tokens are truncated, and the final hidden states are mean-pooled over the attention mask to give $h_t \in \mathbb{R}^{768}$.

Layout. Each hOCR word box $b_i = (x_0, y_0, x_1, y_1)$ is normalised by the page width and height and expanded to $\tilde{b}_i = (x_0, y_0, x_1, y_1, w, h)$, following [3] (Fig. 2a). The word boxes of a page are pooled into a 16×16 text-density map, which is concatenated with six global geometry statistics to form the document-level descriptor $h_\ell \in \mathbb{R}^{262}$ (Fig. 2b). When a word is split into several sub-word tokens, its box is propagated to every token; this alignment is required by a token-level head but, for the document-level descriptor used here, word boxes alone determine h_ℓ . Because CLIP at 224×224 pixels already sees coarse page structure, the contribution of h_ℓ measures what explicit OCR geometry adds beyond that view.

3.4 Gated fusion

Each stream is mapped by a learned projection f_m (for the layout stream, a small encoder) to a shared space, $z_m = f_m(h_m) \in \mathbb{R}^d$ with $d = 256$ and $m \in \{v, t, \ell\}$. A scalar gate is computed for each modality and the modalities are combined as

$$g_m = \sigma(w_m^\top z_m + c_m), \quad \hat{g}_m = \frac{g_m}{\sum_{m'} g_{m'}}, \quad z = \sum_m \hat{g}_m z_m, \quad (1)$$

with $w_m \in \mathbb{R}^d$ and $c_m \in \mathbb{R}$. Each gate is conditioned only on its own modality, whereas in GMU every gate sees all modalities [29]. The class scores are $s = Wz + b$ with $W \in \mathbb{R}^{C \times d}$, followed by a softmax and trained with cross-entropy. Only the projections, gates and linear head are trained; the two frozen encoders contain about 2×10^8 parameters.

3.5 Exact modality attribution

Because the head is linear and z is a weighted sum, each class logit decomposes exactly as

$$s_c = \sum_m \phi_{m,c} + b_c, \quad \phi_{m,c} = \hat{g}_m w_c^\top z_m, \quad (2)$$

where w_c is the c -th row of W . For a given input, $\phi_{m,c}$ is the exact share of modality m in the logit of class c , so this attribution is faithful to the computation by construction and is available for every decision at no extra cost. It differs from the gate weight $\hat{g}_m$ because it also reflects the norm of z_m and its alignment with w_c : a modality with a small

¹<https://huggingface.co/datasets/anirudh1112/corrected-tobacco-dataset-with-ocr>

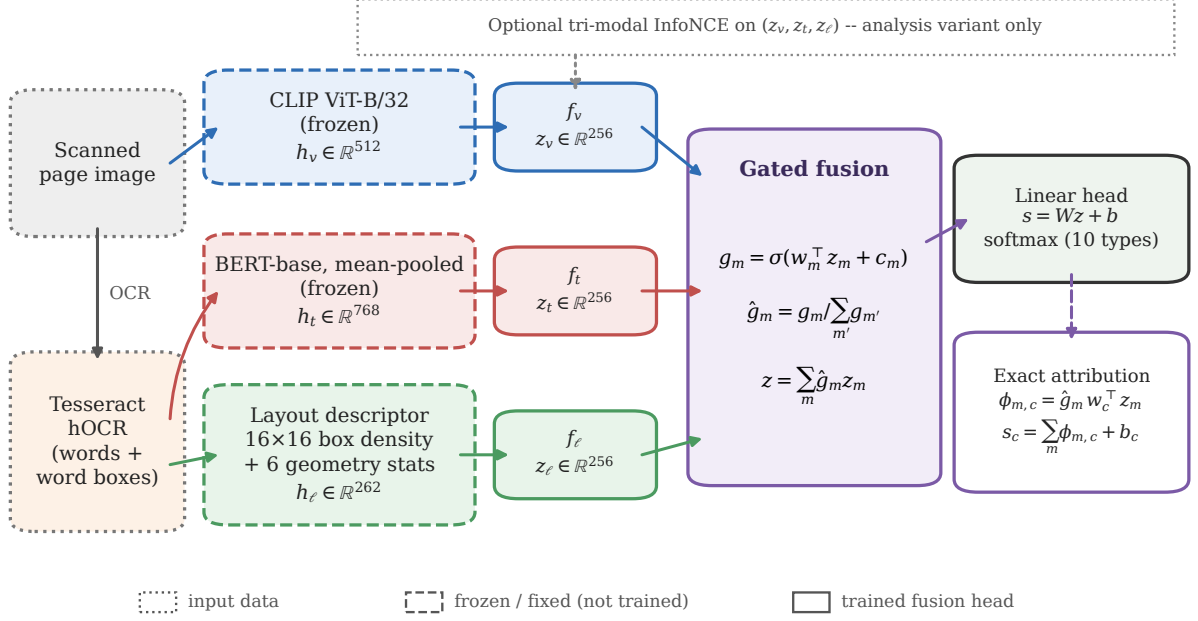


Fig. 1 Overview of GLaDViT. Frozen CLIP and BERT encoders and a fixed layout descriptor computed from hOCR word boxes give three feature vectors, which are projected to a shared 256-dimensional space and combined by a normalised gate. Only the projections, gates and linear head are trained. Because the head is linear, each class logit splits exactly into per-modality terms $\phi_{m,c}$. The triple InfoNCE objective is an optional variant analysed in Section 4.4.

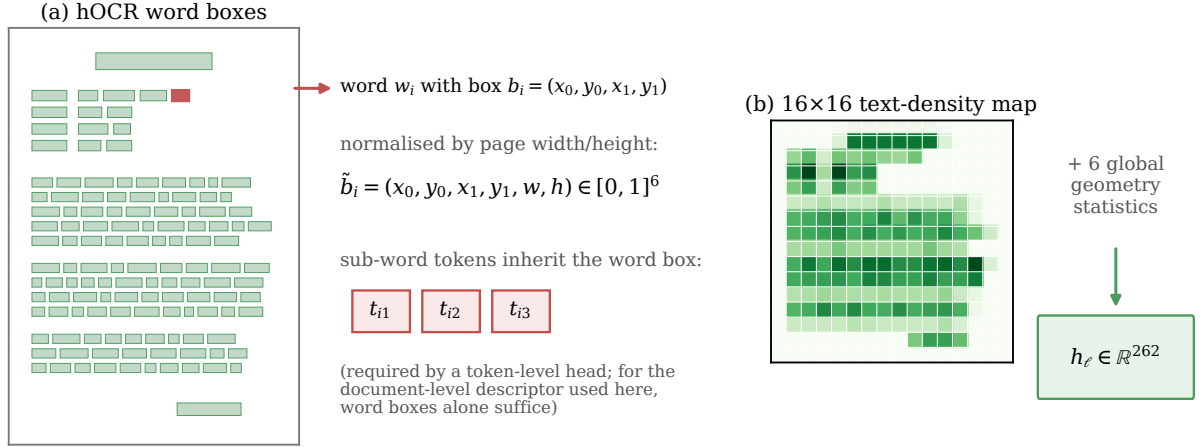


Fig. 2 Layout descriptor. (a) Each hOCR word box is normalised to a six-dimensional vector; sub-word tokens inherit the box of their word. (b) The word boxes of a page are pooled into a 16×16 text-density map (computed here from the boxes drawn in (a)), which is concatenated with six global geometry statistics.

gate but a large, well-aligned embedding can dominate a logit, and vice versa. Equation (2) holds the gates fixed and therefore measures the direct contribution of each modality; the indirect effect that a modality exerts through the normalisation

of the other gates is captured only by total-effect measures such as modality deletion or Shapley values [36, 38].

3.6 Triple contrastive alignment (analysis variant)

For a batch of B documents and a modality pair (a, b) , the InfoNCE loss [31] with learnable temperature τ is

$$\mathcal{L}_{ab} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\langle z_a^i, z_b^i \rangle / \tau)}{\sum_{j=1}^B \exp(\langle z_a^i, z_b^j \rangle / \tau)}, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ is cosine similarity. The three pairs are combined as

$$\mathcal{L}_{\text{con}} = \lambda_{vt} \mathcal{L}_{vt} + \lambda_{tl} \mathcal{L}_{tl} + \lambda_{vl} \mathcal{L}_{vl}, \quad (4)$$

with weights $\lambda_{vt}, \lambda_{tl}, \lambda_{vl}$ (dashed box in Fig. 1). In the variant evaluated here, the projection heads are first warmed up with Eq. (4) and the model is then trained with cross-entropy.

3.7 Training and evaluation

Features are precomputed once. Training uses Adam [39] (learning rate 10^{-3} , weight decay 10^{-4}) with early stopping on validation macro-F1. The gated model is compared with single-modality heads and with mean fusion (the unweighted average of the projected modalities) on the same features and split. Results are the mean $\pm$ standard deviation over three seeds. We report accuracy, macro-F1 and weighted-F1, the test confusion matrix and the mean gate weight $\hat{g}_m$ per class. For test accuracy we give the Wilson 95% interval [40]. With 279 test pages, one page equals 0.36 accuracy points, so a difference of about one point involves only two or three pages; confirming such a difference requires paired tests on per-page predictions, such as McNemar’s test [41, 42], and we interpret small differences accordingly.

4 Results

4.1 Ablation

Table 1 reports the ablation on the corrected release. Among single modalities, vision is strongest (87.7% accuracy), followed by text (84.0%) and layout (60.2%); layout is weak alone but complementary. Combining vision and text by mean fusion gives 89.8%; adding the layout stream raises accuracy by 2.8 points (about eight

test pages), and replacing mean fusion by the gate adds a further 1.1 points (1.0 weighted-F1 point, about three test pages), for a best result of $93.7 \pm 0.4\%$ accuracy and 93.6% weighted F1. The Wilson intervals of the mean-fusion and gated models overlap substantially, so the gating gain should be regarded as indicative; the layout gain is larger (about eight pages) but likewise requires paired testing.

Fig. 3 shows the ablation and the test confusion matrix of the gated model. The matrix is strongly diagonal; recall is highest for memo, note and report, and the main confusions involve email, which is mistaken for form, report, news or resume, and form, which is mistaken for letter.

4.2 Comparison with the literature

Table 2 lists published Tobacco-3482 results together with the modalities, whether the model was first trained on RVL-CDIP, and the label version. Accuracies reported on RVL-CDIP (for example 92.1% for DiT-B [43] and 95.9% for LayoutLMv3-large [5]) concern a different dataset and are not included. Among methods without RVL-CDIP transfer, the best published result we found is 91.14% (WordVIS [17]). GLaDViT obtains 93.67% on the corrected release; because the labels and split differ, this value is indicative and not a head-to-head comparison. Methods with RVL-CDIP transfer (up to 95.57% [7]) benefit from in-domain pretraining on the same IIT-CDIP collection and form a separate category.

4.3 Interpretability

Gate weights. Fig. 4 shows the mean normalised gate weights $\hat{g}_m$ per class on the test set. They follow document intuition: visually distinctive types put most gate mass on vision (news 89%, advertisement 81%, letter 80%), whereas text-dense types shift mass to text (memo 65%; report, resume and scientific about 40%), and forms and notes receive the largest layout share (about 10%).

Gate weights are not importance scores.

The layout stream receives at most about 10% of the gate mass, and almost none for letters and emails, yet adding it to mean-fused vision and text improves accuracy by 2.8 points. Reading the gate weights as explanations, as is common

Table 1 Results on the corrected Tobacco-3482 release (train = 2186, val = 272, test = 279), frozen backbones, mean $\pm$ standard deviation over three seeds. The Wilson 95% interval is computed at the seed-mean accuracy with $n = 279$.

Configuration	Acc. (%)	95% CI	Macro F1 (%)	Weighted F1 (%)
Vision only	87.7 ± 0.3	83.3–91.0	88.0 ± 0.4	87.5 ± 0.4
Text only	84.0 ± 0.2	79.2–87.8	82.1 ± 0.0	84.0 ± 0.2
Layout only (hOCR)	60.2 ± 0.8	54.4–65.8	55.4 ± 1.6	59.1 ± 0.9
V + T, mean fusion	89.8 ± 1.0	85.7–92.8	90.9 ± 0.8	89.8 ± 1.0
V + T + L, mean fusion	92.6 ± 0.2	88.9–95.1	92.4 ± 0.5	92.6 ± 0.2
V + T + L, gated (GLaDViT)	93.7 ± 0.4	90.2–96.0	93.7 ± 0.5	93.6 ± 0.4
+ triple contrastive warm-up	90.3 ± 0.8	86.3–93.2	90.8 ± 0.7	90.2 ± 0.8

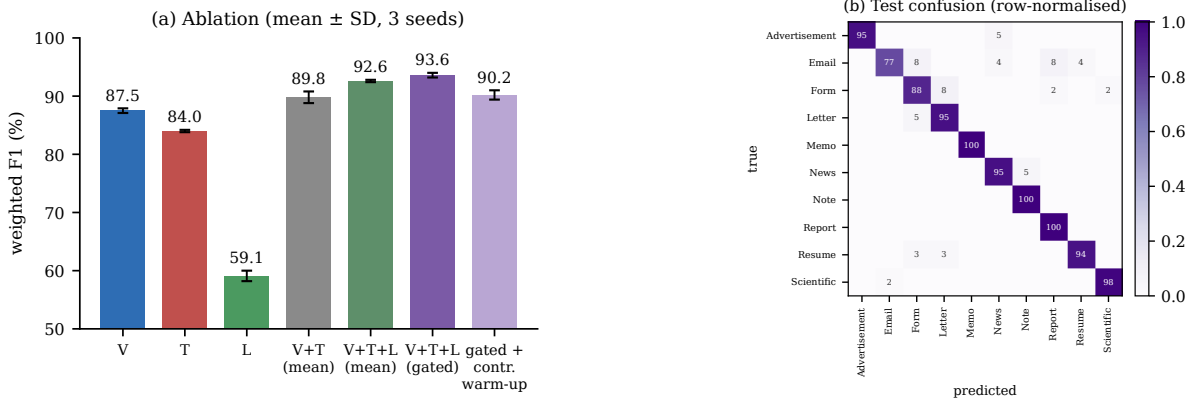


Fig. 3 Results on the corrected Tobacco-3482 release. (a) Weighted F1 of the ablation (mean of three seeds; error bars are standard deviations; the y -axis starts at 50%). (b) Row-normalised test confusion matrix of the gated model (values in %).

for attention-like mechanisms [33], would therefore understate the role of layout. Equation (2) explains the gap: the share of a modality in a logit depends on $\hat{g}_m$ and also on the norm of z_m and its alignment with the class weight w_c , so a stream with a small gate can still shift the decision. The decomposition $\phi_{m,c}$ gives this share exactly for every prediction and is therefore the attribution we recommend reporting alongside, or instead of, the gate weights.

4.4 Triple contrastive alignment

Warming up the projection heads with the triple contrastive objective lowers accuracy from 93.7% to $90.3 \pm 0.8\%$ (weighted F1 90.2%; Table 1), a loss of about ten test pages that lies outside the seed variation of either configuration.

5 Discussion

The results support three claims. First, the three streams are complementary: the best single modality reaches 87.7% and the three-stream model 93.7%, and layout, although weak alone, contributes about eight test pages when added to vision and text. Second, gating adds a smaller gain over mean fusion, about three test pages, which is within the overlap of the confidence intervals and should be read as indicative until confirmed by paired tests. Third, gate weights and modality importance diverge: the layout stream carries little gate mass but a measurable effect, and the exact decomposition of Eq. (2) provides a faithful account of each decision at no extra inference cost.

The triple contrastive objective did not help in the frozen-backbone setting. Three explanations are plausible: because the same projection heads feed both the alignment loss and the gate, the

Table 2 Published Tobacco-3482 accuracies as reported in the cited works. RVL: model first trained on RVL-CDIP. Earlier results use the original labels and their own train/test protocols, most following Kang et al. [16]; GLaDViT uses the corrected labels and fixed split of Section 3.2. Values in different groups are therefore indicative rather than head-to-head.

Method	Modality	Labels	RVL	Acc. (%)
AlexNet, ImageNet init. [6]	image	original	no	75.73
VGG-16, ImageNet init. [6]	image	original	no	77.52
EfficientNet [19]	image	original	no	85.99
Audebert et al. [18]	image+text	original	no	87.80
EfficientNet + BERT [19]	image+text	original	no	89.47
DocXClassifier [7]	image	original	no	90.14
WordVIS + DocXClassifier [17]	image (text rendered)	original	no	91.14
ResNet-50, RVL-CDIP init. [6]	image	original	yes	91.13
EfficientNet, RVL-CDIP init. [19]	image	original	yes	94.04
DocXClassifier-L, RVL-CDIP init. [7]	image	original	yes	95.57
GLaDViT (ours)	image+text+layout	corrected	no	93.67

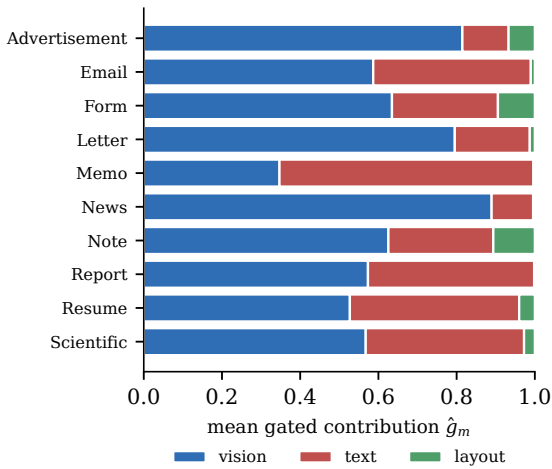


Fig. 4 Mean normalised gate weights $\hat{g}_m$ per class on the Tobacco-3482 test set. They are mixing coefficients rather than importance scores (Section 4.3).

alignment pulls them towards cross-modal agreement and away from class-discriminative directions; small batches provide few negatives; and frozen features leave only a low-capacity projection to align. Joint training with a down-weighted alignment term, separate contrastive heads discarded after warm-up as in SimCLR [44], and a low-label regime would test these explanations. Contrastive alignment is expected to pay off when the encoders themselves adapt [14].

Several limitations remain. The results use a single fixed split of a label-corrected subset whose remaining pages are less ambiguous than the original set, and a test set of 279 pages gives confidence intervals of about ± 3 points; an evaluation under

the classic protocol of Kang et al. [16] on the original labels, paired significance tests, and further fusion baselines on the same features (concatenation, late fusion, static weights and GMU [29]) are needed to place the results more firmly. The attribution analysis reports gate weights and the exact direct decomposition; total-effect measures such as modality deletion and Shapley values [36] remain to be computed. The evaluation uses one English dataset and one OCR engine at fixed resolution, and OCR quality and latency directly affect the text and layout streams. The layout stream is a document-level descriptor rather than a per-token encoder, and the token-level extension for key information extraction is not evaluated.

6 Conclusion

We presented GLaDViT, a gated fusion of frozen CLIP, BERT and hOCR-layout features for document image classification in which only a small head is trained. On a label-corrected release of Tobacco-3482 it reaches 93.7% accuracy (95% interval 90.2–96.0%). Layout contributes beyond image and text, and an exact logit decomposition provides faithful per-decision modality attributions that differ from the gate weights themselves. A triple contrastive objective does not help when the backbones are frozen. Future work will evaluate the model under the classic protocol with paired significance tests and additional fusion

baselines, fine-tune the encoders end to end, evaluate a token-level head for key information extraction, and extend the evaluation to RVL-CDIP and to other scan qualities and languages.

Acknowledgements. The author thanks İS Teknoloji for the collaborative research environment in which this work was initiated.

Declarations

- **Funding:** This study received no external funding.
- **Competing interests:** The author declares no competing interests.
- **Ethics approval and consent to participate:** Not applicable. The study uses only a publicly available dataset of scanned documents.
- **Consent for publication:** Not applicable.
- **Data availability:** Tobacco-3482 is publicly available [2]. The label-corrected release used in this study is available at <https://huggingface.co/datasets/anirudh1112/corrected-tobacco-dataset-with-ocr>.
- **Materials availability:** Not applicable.
- **Code availability:** Code, precomputed features and the hOCR token-box tooling will be deposited in a public repository with a permanent DOI upon acceptance; during review they are available from the corresponding author on reasonable request.
- **Author contribution:** M.Ş. is the sole author and was responsible for conceptualisation, methodology, software, experiments, analysis, visualisation and writing.

References

- [1] Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp. 991–995 (2015)
- [2] Kumar, J., Ye, P., Doermann, D.: Structural similarity for document image classification and retrieval. *Pattern Recognition Letters* **43**, 119–126 (2014)
- [3] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: pre-training of text and layout for document image understanding. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1192–1200 (2020)
- [4] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., *et al.*: LayoutLMv2: multi-modal pre-training for visually-rich document understanding. In: Proceedings of ACL-IJCNLP, pp. 2579–2591 (2021)
- [5] Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: pre-training for document AI with unified text and image masking. In: Proceedings of the ACM International Conference on Multimedia, pp. 4083–4091 (2022)
- [6] Afzal, M.Z., Kölsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp. 883–888 (2017)
- [7] Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: DocXClassifier: towards a robust and interpretable deep neural network for document image classification. *International Journal on Document Analysis and Recognition* **27**, 447–473 (2024)
- [8] Smith, R.: An overview of the Tesseract OCR engine. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), pp. 629–633 (2007)
- [9] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., *et al.*: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)
- [10] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
- [11] Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should I trust you?”: explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
- [12] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
- [13] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., *et al.*: ImageBind: one embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180–15190 (2023)

- [14] Bakkali, S., Ming, Z., Coustaty, M., Rusiñol, M., Ramos Terrades, O.: VLCDoC: vision-language contrastive pre-training model for cross-modal document classification. *Pattern Recognition* **139**, 109419 (2023)
- [15] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., *et al.*: Building a test collection for complex document information processing. In: *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 665–666 (2006)
- [16] Kang, L., Kumar, J., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for document image classification. In: *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 3168–3172 (2014)
- [17] Khan, U., Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: WordVIS: a color worth a thousand words. *arXiv:2412.10155* (2024)
- [18] Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. *Communications in Computer and Information Science* **1167**, 427–443 (2020)
- [19] Ferrando, J., Domínguez, J.L., Torres, J., García, R., García, D., *et al.*: Improving accuracy and speeding up document image classification through parallel systems. In: *Computational Science – ICCS 2020. Lecture Notes in Computer Science*, vol. 12138, pp. 387–400 (2020)
- [20] Lim, G., Larson, S., Leach, K.: Label errors in the Tobacco3482 dataset. *arXiv:2412.13140* (2024)
- [21] Larson, S., Lim, G., Leach, K.: On evaluation of document classification with RVL-CDIP. In: *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (2023)
- [22] Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: DocFormer: end-to-end transformer for document understanding. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 993–1003 (2021)
- [23] Wang, J., Jin, L., Ding, K.: LiLT: a simple yet effective language-independent layout transformer for structured document understanding. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7747–7757 (2022)
- [24] Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., *et al.*: Unifying vision, text, and layout for universal document processing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19254–19264 (2023)
- [25] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., *et al.*: OCR-free document understanding transformer. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 498–517 (2022)
- [26] Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., *et al.*: Pix2Struct: screenshot parsing as pretraining for visual language understanding. In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 18893–18912 (2023)
- [27] Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., *et al.*: mPLUG-DocOwl 1.5: unified structure learning for OCR-free document understanding. In: *Findings of the Association for Computational Linguistics: EMNLP* (2024)
- [28] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P.: Tensor fusion network for multimodal sentiment analysis. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1103–1114 (2017)
- [29] Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal units for information fusion. *ICLR Workshop*, *arXiv:1702.01992* (2017)
- [30] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2), 423–443 (2019)
- [31] Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv:1807.03748* (2018)
- [32] Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., *et al.*: Contrastive multimodal fusion with TupleInfoNCE. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 754–763 (2021)
- [33] Jain, S., Wallace, B.C.: Attention is not explanation. In: *Proceedings of NAACL-HLT*, pp. 3543–3556 (2019)
- [34] Wiegrefe, S., Pinter, Y.: Attention is not not explanation. In: *Proceedings of EMNLP-IJCNLP*, pp. 11–20 (2019)
- [35] Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4198–4205 (2020)
- [36] Parcalabescu, L., Frank, A.: MM-SHAP: a performance-agnostic metric for measuring multimodal contributions in vision and language models and tasks. In: *Proceedings of the Annual*

- Meeting of the Association for Computational Linguistics (ACL) (2023)
- [37] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., *et al.*: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
 - [38] Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games II, pp. 307–317. Princeton University Press, Princeton, NJ (1953)
 - [39] Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
 - [40] Wilson, E.B.: Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**(158), 209–212 (1927)
 - [41] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157 (1947)
 - [42] Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation* **10**(7), 1895–1923 (1998)
 - [43] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F.: DiT: self-supervised pre-training for document image transformer. In: Proceedings of the ACM International Conference on Multimedia, pp. 3530–3539 (2022)
 - [44] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1597–1607 (2020)