Explaining Feature Contributions in Multimodal Instagram Engagement Prediction

Carolina Menconi[1] and Laura Pollacci[1](✉)

Department of Computer Science, University of Pisa, Italy,
laura.pollacci@unipi.it

Abstract. Engagement is an observable outcome for analyzing content diffusion in image-based social media. This paper presents MIXER, an integrated protocol that compares multimodal representations and explains feature contributions in engagement prediction from aligned metadata, captions, and images. We evaluate MIXER on a large Instagram dataset with follower-normalized labels at three granularities. We compare traditional, semantic, and vision–language representations in unimodal, multimodal, and CLIP-based fusion settings. Multimodal models improve over unimodal alternatives in most settings. Macro-F1 is lower for finer-grained tasks, although scores across different numbers of classes are interpreted descriptively rather than as directly comparable performance estimates. Post-hoc XAI shows that text contributes least, image and metadata account for most attributions, and metadata is especially relevant in the five-class setting and at the extremes. MIXER clarifies how modalities and feature groups contribute to engagement prediction across target granularities and representation settings.

Keywords: Instagram engagement prediction · multimodal learning · post-hoc XAI · feature attribution · content diffusion

1 Introduction

Instagram engagement provides a measurable outcome for analyzing content diffusion in platform-mediated communication. Likes, comments, and related interactions do not measure ranking, reach, or recommendation mechanisms, but provide a quantitative view of user response at scale. Identifying post- and account-level features associated with engagement is therefore relevant to analyses of online visibility and information environments.

Engagement prediction is intrinsically multimodal because Instagram posts combine images, captions, hashtags, temporal information, and account metadata. Prior work shows that visual, textual, and metadata features support popularity or engagement prediction [1,5,6,8–11]. Predictive performance alone does not reveal the relative role of post content, metadata, and account structure.

We present the MultImodal eXplainable Engagement fRamework (MIXER)[1], an integrated protocol for multimodal engagement prediction and post-hoc ex-

[1] Code available at: https://github.com/LauraPollacci/MIXER.

planation. It compares traditional, semantic, and vision–language representations in unimodal and multimodal configurations. We evaluate it on the Instagram Influencers Dataset [4], using follower-normalized binary, three-class, and five-class labels, with macro-F1 as the primary metric. A reduced-subset experiment compares CLIP-based concatenation and cross-attention.

MIXER addresses four questions: *Q1)* whether multimodal models improve over text-only and image-only alternatives; *Q2)* which representation family performs best under the same target definitions; *Q3)* whether cross-attention improves over concatenation in the reduced-subset CLIP setting; and *Q4)* how results change across the three granularities. The explainability analysis assesses modality and feature-group contributions within the selected configuration.

Multimodal configurations are generally stronger than unimodal alternatives. Macro-F1 is lower in tasks with more classes, but cross-granularity values are interpreted descriptively because the number of classes and chance level change. Text contributes less than image and metadata, while account-level structural features become especially relevant in the five-class setting and at the extremes.

The paper contributes: *(i)* MIXER, an integrated protocol for multimodal engagement prediction and post-hoc explanation; *(ii)* a comparison of three representation families across binary, three-class, and five-class tasks; *(iii)* a reduced-subset comparison of CLIP-based concatenation and cross-attention; and *(iv)* global and local attribution analyses of modality and feature-group contributions. Instagram provides a large-scale image-based communication environment where visual content, text, and account information are jointly available. Engagement is treated as an observable outcome associated with content circulation in platform-mediated public communication. The analysis does not measure democratic outcomes, exposure, or algorithmic amplification, but addresses one observable component of online information environments.

The rest of the paper is organized as follows. Section 2 reviews related work, Section 3 introduces MIXER, and Section 4 describes the empirical setting. Sections 5 and 6 report predictive and explainability results. Section 7 discusses implications and limitations, and Section 8 concludes the paper.

2 Related Work and Positioning

Comparisons across social-media engagement studies are difficult because platforms, measures, tasks, features, splits, and architectures vary. In Instagram research [1, 5, 6, 8–11], engagement or popularity is commonly based on likes, comments, or normalized variants and formulated as regression, ranking, or classification. This heterogeneity motivates comparisons across representations, modalities, fusion strategies, and label granularities. Early Instagram popularity studies combined visual, textual, and contextual features. The work in [5] frames brand-related posts as a learning-to-rank problem based on likes, while [6] shows that category-specific models improve ranking. Other contributions focus on specific accounts, feature groups, or modalities. Visual deep-learning models with Grad-CAM explanations are used in [11]; post and account metadata support the

detection of deviations from recent account-level performance in [1]; and [8] proposes an explainable baseline based mainly on visual and user-level information, using SHAP for feature-group analysis. These works show that heterogeneous feature groups are informative, yet their contributions are only partly analyzed.

Recent work increasingly treats interpretability as a design requirement. The approach in [10] combines handcrafted visual, textual, and metadata features with decision trees that provide human-readable rules. The work in [9] proposes an evidential reasoning approach for Instagram post popularity and compares it with standard machine-learning classifiers. Interpretability is mainly provided through interpretable models, handcrafted features, or local visual explanations. Post-hoc XAI remains less explored in multimodal engagement prediction, especially for comparing feature contributions across modalities and engagement granularities. Evidence from influencer marketing further supports combining image, text, and influencer-level information. The authors in [3] find that multimodal information improves Instagram engagement prediction and that image and influencer features contribute more than text. Our results confirm these patterns while extending the comparison across representation families, label granularities, and global and local explanation levels.

Engagement measures also vary. Some works use absolute likes [5,6,8,9], while others normalize engagement by followers to reduce the direct effect of audience size across heterogeneous accounts [10–12]. Following [2], we use an engagement rate based on likes and comments normalized by followers. This reduces the target's direct dependence on audience size, although account-level effects may remain. Label granularity differs as well. Several works formulate engagement prediction as a binary task [1,9–11], while others use multi-class formulations [11, 12]. We include binary, three-class, and five-class targets to assess how predictive performance and feature contributions change across engagement levels.

Prior work shows that combining image, text, metadata, and account-level features supports Instagram engagement prediction, although the task remains difficult because engagement depends on post content and account structure. Explainability is often limited to interpretable model classes, handcrafted features, or local visual explanations. MIXER builds on this work by comparing representation families, fusion strategies, and label granularities, with global and local post-hoc XAI for modality- and feature-group contributions.

3 The MIXER Framework

MIXER is a methodological framework for multimodal engagement prediction and post-hoc explanation of feature contributions in image-based social media posts. Figure 1 summarizes the MIXER pipeline. Data preparation defines the target, text/image/metadata representations are extracted, unimodal and multimodal models are trained and evaluated, and global/local XAI is applied to the selected model. In this paper, MIXER is evaluated on Instagram posts, represented through three sources of information: structured metadata, caption text, and visual content. This design supports two connected analyses: *a)* measuring

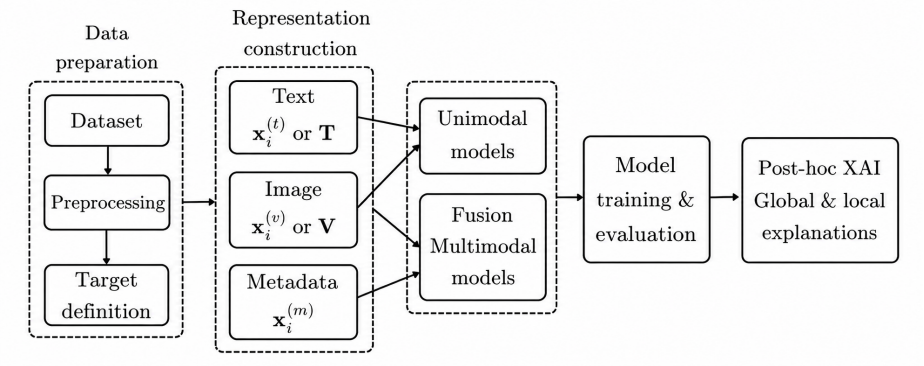


Fig. 1: MIXER Framework.

how well engagement levels can be predicted, and *b)* inspecting how modalities and feature groups contribute across engagement levels.

We formalize engagement prediction over Instagram posts $\mathcal{P} = \{p_1, \ldots, p_N\}$. Each post p_i has three aligned views: metadata $\mathbf{x}_i^{(m)}$, text $\mathbf{x}_i^{(t)}$, and image $\mathbf{x}_i^{(v)}$. The target is an ordered class $y_i \in \mathcal{Y}$ obtained by discretizing a log-transformed, follower-normalized engagement rate:

$$ER_i^{(\log)} = \log\left(1 + 100 \cdot \frac{likes_i + comments_i}{followers_i}\right) \tag{1}$$

Here, $likes_i$ and $comments_i$ are the reactions to post p_i, $followers_i$ is the follower count, 100 expresses a percentage, and the additive constant handles zero-valued rates. The transformed target is discretized by quantiles: the median defines the binary setting, while additional thresholds define the three-class and five-class labels. The resulting classes retain an ordinal interpretation, but the current protocol treats them as nominal during model training and macro-F1 evaluation. The three granularities are therefore used to compare attribution patterns and to report descriptive changes in predictive performance, rather than to provide directly normalized estimates of task difficulty.

3.1 Multimodal Representations

MIXER represents each post through metadata, text, and image features.

The *metadata view* captures structured information about the post and the account. It includes temporal features, caption-structure features, image geometry, creator category, and account-level structural features. Categorical metadata variables are encoded as numerical indicators, while numerical metadata variables are standardized. The resulting metadata representation is denoted by $\mathbf{x}_i^{(m)} \in \mathbb{R}^{d_m}$, where d_m is the number of metadata features.

The *textual view* represents the caption. MIXER supports sparse lexical representations and dense semantic embeddings. Sparse features use TF–IDF vectors enriched with category-specific lexical indicators, while dense features use

pretrained language or vision–language encoders. The resulting textual representation is denoted by $\mathbf{x}_i^{(t)} \in \mathbb{R}^{d_t}$, where d_t is its dimensionality. For token-level fusion, the caption can also be represented as $\mathbf{T}_i \in \mathbb{R}^{L_t \times d}$, where L_t is the number of text tokens and d is the token-embedding dimensionality.

The *visual view* represents the image content of the post. MIXER supports handcrafted image descriptors, dense visual embeddings, and vision–language image embeddings. For posts containing more than one image, image-level representations are aggregated into a single post-level vector. The resulting visual representation is denoted by $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$, where d_v is the dimensionality of the visual representation. For token-level or patch-level fusion, the image can also be represented as a sequence $\mathbf{V}_i \in \mathbb{R}^{L_v \times d}$, where L_v is the number of visual tokens or image patches and d is the corresponding embedding dimensionality.

3.2 Fusion and Prediction

MIXER compares *unimodal* configurations, based on one source of information, with *multimodal* configurations, which combine modalities into joint representations. The main multimodal strategy is *early fusion by concatenation*: $\mathbf{x}_i^{(\text{multi})} = \mathbf{x}_i^{(m)} \oplus \mathbf{x}_i^{(t)} \oplus \mathbf{x}_i^{(v)}$, where $\oplus$ denotes vector concatenation. This model-agnostic strategy combines sparse lexical features, dense embeddings, handcrafted descriptors, and structured metadata.

In the vision–language setting, MIXER also compares concatenation with two cross-attention mechanisms. Let $\mathbf{T}_i \in \mathbb{R}^{L_t \times d}$ be the sequence of text-token embeddings and $\mathbf{V}_i \in \mathbb{R}^{L_v \times d}$ the sequence of image-patch embeddings. The two cross-attention variants are *direct cross-attention* on frozen CLIP representations and *fine-tuned cross-attention* with task-specific training. Both variants let one modality condition its representation on the other before producing a pooled multimodal representation. This comparison tests whether more expressive fusion improves over concatenation in the CLIP-based setting.

The resulting representation is passed to a classifier that predicts the engagement class, $\hat{y}_i = f(\mathbf{z}_i)$, where $\mathbf{z}_i$ denotes either a unimodal or fused representation, f is the trained classifier, and $\hat{y}_i$ is the predicted label for post p_i. The framework is classifier-agnostic; here, we compare linear, probabilistic, ensemble, and gradient-boosted classifiers.

3.3 Explainability Protocol

MIXER uses post-hoc XAI to explain feature contributions after model selection. We apply *global* and *local* explanations to the selected multimodal configuration to assess how modalities and feature groups contribute within the same predictive model. Global explanations estimate which modalities and feature groups contribute most to the predictions, while local explanations perturb image regions or caption spans and measure the resulting change in predicted probability.

For global explanations, we use SHAP values. Let $\phi_{ij}^{(c)}$ be the SHAP attribution of feature j for post p_i and class c. In multiclass settings, class-specific

attributions are summarized as $\bar{\phi}_{ij} = C^{-1} \sum_{c=1}^{C} |\phi_{ij}^{(c)}|$, where $C = |\mathcal{Y}|$. Global feature importance is then obtained by averaging $\bar{\phi}_{ij}$ over posts. View-level importance is computed by summing feature importances within the metadata, text, and image views. The resulting shares measure the aggregate contribution of each view to the model; they are not normalized by the number of features in the view and should not be read as average per-feature scores. Because individual dimensions of dense text and image embeddings do not have a direct semantic interpretation, feature-level inspection is restricted to metadata variables.

For local explanations, MIXER uses perturbation-based analysis. For a post p_i, let $\hat{c}$ be the predicted class. Given a perturbed version p_i', the signed perturbation effect is $\Delta_i = P(\hat{c} \mid p_i) - P(\hat{c} \mid p_i')$. Here, p_i' is obtained by masking an image region or removing a caption token or span. A positive value indicates that the perturbed component supported the original prediction, because removing it decreases the probability of $\hat{c}$; a negative value indicates the opposite effect. For images, the post image is divided into a $g \times g$ grid and one cell is masked at a time with a neutral value. For each masked image, the image representation is recomputed and Δ_i is measured, producing a grid-level attribution map. For captions, one token or a contiguous span of tokens is removed, the caption representation is recomputed, and the same signed probability change is measured. These local explanations inspect model sensitivity under controlled input perturbations; they are not interpreted as causal effects.

4 Experimental Setting

This section describes the dataset, target construction, representations, training protocol, and evaluation questions. Appendices A, B, and C provide details on preprocessing, feature extraction, and model configurations.

4.1 Dataset and Target Construction

MIXER is evaluated on the Instagram Influencers Dataset [4], which contains 33,935 influencers from nine categories: beauty, family, fashion, fitness, food, interior, pet, travel, and other. It includes up to 300 posts per influencer, for approximately 10.18 million posts, each with metadata, caption text, engagement variables, and images. The target is derived from the log-transformed engagement rate in Equation 1, whose distribution is strongly right-skewed with a long upper tail. Quantile-based definitions produce approximately balanced classes while preserving their ordinal structure. We consider three granularities: the binary setting uses the median to define `low` and `high`; the three-class setting uses the 0.33 and 0.66 quantiles; and the five-class setting uses the 0.2, 0.4, 0.6, and 0.8 quantiles. Table 1 shows the corresponding ranges on the original ER scale. Their uneven width, especially in the upper classes, reflects the long-tailed distribution and produces broader intervals at higher engagement. Because labels are quantile-based rather than raw regression targets, extreme engagement rates widen the upper range without dominating the resulting classes.

Table 1: ER discretization.

Setting	Class	Min	Max	Avg
Binary	`low`	0.00	3.04	1.61
Binary	`high`	3.05	1409.59	7.13
3-class	`low`	0.00	2.00	1.13
3-class	`medium`	2.01	4.45	3.09
3-class	`high`	4.46	1409.59	8.83
5-class	`very_low`	0.00	1.30	0.77
5-class	`low`	1.31	2.40	1.84
5-class	`medium`	2.41	3.86	3.08
5-class	`high`	3.87	6.46	5.00
5-class	`very_high`	6.47	1409.59	11.40

Table 2: Class distribution (%).

Setting	Class	Train	Val.	Test
Binary	`low`	49.91	50.11	49.87
Binary	`high`	50.09	49.89	50.13
3-class	`low`	31.88	32.81	33.02
3-class	`medium`	35.54	33.27	32.72
3-class	`high`	32.57	33.91	34.26
5-class	`very_low`	17.92	19.47	19.51
5-class	`low`	21.51	20.71	20.58
5-class	`medium`	21.59	19.98	19.80
5-class	`high`	21.47	20.12	19.51
5-class	`very_high`	17.51	19.71	20.61

The analysis includes 2017–2018 posts with English captions and at least one valid image. The language restriction reduces textual heterogeneity, while the image constraint ensures a multimodal representation for each post. For posts with multiple images, image-level representations are aggregated into one post-level visual vector. Images are converted to RGB and resized to 224×224; captions are cleaned by removing hyperlinks, email addresses, and user mentions, while hashtags are reduced to their textual component. After preprocessing and training-only category balancing (Section 4.2), the train, validation, and test splits contain 773,497, 412,325, and 423,604 posts, with 960,048, 556,982, and 588,557 images, respectively. Further details are provided in Appendix A.

4.2 Temporal Split, Balancing, and Evaluation Protocol

The split is chronological at post level: training covers January 2017–October 2018, validation November 2018, and test December 2018. The intended scenario is temporal prediction of engagement classes for later posts by observed creators. The protocol does not enforce a creator-disjoint split and therefore does not evaluate generalization to previously unseen creators.

Category imbalance is addressed only in training. The smallest category, *pet*, contains 544 influencers and defines the number selected from each category. All *pet* influencers are retained; for each other category, 544 are selected using sorted username hashes, and all their training posts are retained. Validation and test remain unchanged and preserve their natural distributions. This procedure reduces majority-category dominance during model fitting, but the current experiments do not include an ablation measuring its effect on test performance.

The quantile-based discretization produces approximately balanced engagement classes, although their proportions vary slightly across splits because of the chronological separation and training-only category balancing. Table 2 reports these distributions. Macro-F1 is used as the primary metric because it gives equal weight to all classes within each target formulation. Additional details on the split and balancing protocol are provided in Appendix A.

4.3 Representations and Model Configurations

The experiments compare three representation families, each combining text and image representations with metadata. The *traditional* family uses TF–IDF caption features and 149 handcrafted visual descriptors; the *semantic* family uses Sentence-BERT (SBERT) caption embeddings and EfficientNet-B0 image embeddings; and the *vision–language* family uses Contrastive Language–Image Pre-training (CLIP) [7], which maps captions and images into aligned embedding spaces. Multimodal representations are aligned at post level and concatenated with metadata, which includes temporal, caption-structure, image-geometry, creator-category, and account-level structural features.

Identifiers and variables used to compute the target, including likes, comments, and followers, are excluded from prediction. The validation-selected TF–IDF setting uses unigrams, `min_df`= 20, `max_df`= 0.7, up to 50,000 features, and category-specific top-word indicators. SBERT, EfficientNet-B0, and CLIP produce 384-, 1280-, and 512-dimensional embeddings, respectively. For posts with multiple images, image embeddings are mean-pooled into one post-level vector. Appendix B describes metadata groups and handcrafted visual descriptors. All data-dependent transformations, including TF–IDF vocabularies, encoders, and scalers, are fitted on training data and applied unchanged to validation and test. We compare Linear SVM, Gaussian Naive Bayes, Random Forest, and XGBoost, selecting the classifier and its hyperparameters on validation and reporting performance on the held-out test set. The CLIP experiments compare concatenation and cross-attention on a reduced subset of 25,000 training, 5,000 validation, and 5,000 test posts for computational reasons. Results apply only to this setting. Appendix C provides classifier details.

5 Predictive Results

This section presents the predictive results obtained with MIXER, following the four questions introduced in Section 1. The analysis first compares multimodal representation families, then modality configurations, and finally fusion mechanisms and label granularity.

5.1 Multimodal Performance Across Representation Families

Table 3 reports the main predictive results across representation families, modality settings, and target granularities. In this subsection, we focus on the multimodal rows (*Multimodal* for traditional, SBERT + EfficientNet, and CLIP) to compare the three representation families when text, image, and metadata are combined. This comparison addresses which representation family performs best in the multimodal setting. For these multimodal configurations, XGBoost is selected in all target settings; classifier-level details are reported in Appendix C.

The multimodal rows show essentially aligned observed performance for the traditional and vision–language configurations. Both reach 0.69 binary macro-F1, while the vision–language configuration is 0.01 higher in the three-class and

Table 3: Macro-F1 by modality and representation; bold marks column maxima and underlining within-family maxima.

Family	Config.	Binary	3-class	5-class
Traditional	Text	0.61 (SVM)	0.43 (SVM)	0.29 (SVM)
Traditional	Image	0.57 (XGB)	0.39 (XGB)	0.25 (XGB)
Traditional	Multimodal	**0.69** (XGB)	0.51 (XGB)	0.33 (XGB)
SBERT + EfficientNet	Text	0.58 (XGB)	0.40 (XGB)	0.25 (XGB)
SBERT + EfficientNet	Image	0.59 (SVM/XGB)	0.41 (SVM/RF/XGB)	0.26 (XGB)
SBERT + EfficientNet	Multimodal	0.60 (SVM/XGB)	0.42 (XGB)	0.26 (XGB)
CLIP	Text	0.58 (SVM/XGB)	0.41 (SVM/XGB)	0.25 (SVM/XGB)
CLIP	Image	0.62 (SVM/XGB)	0.44 (XGB)	0.19 (RF)
CLIP	Multimodal	**0.69** (XGB)	**0.52** (XGB)	**0.34** (XGB)

Table 4: CLIP fusion on the reduced subset; maxima are bold.

Fusion	Binary	3-class	5-class
Concat.	**0.68**	**0.50**	**0.33**
Direct cross-att.	0.66	0.45	0.31
Fine-tuned cross-att.	0.62	0.43	0.28

five-class settings. Because the experiments do not include repeated runs or uncertainty estimates, these small differences are not interpreted as evidence that the vision–language representation is superior. By contrast, the semantic configuration is consistently lower across all three target formulations. Across the three separately defined tasks, the highest observed macro-F1 values are 0.69, 0.52, and 0.34. However, the number of classes and the corresponding chance level change across formulations, so these values are not directly comparable as normalized estimates of task difficulty. We therefore report the downward pattern descriptively, without attributing it solely to the finer engagement distinctions.

5.2 Modality and Representation Effects

Within each representation family, Table 3 compares text-only, image-only, and multimodal configurations to assess whether combining views improves over unimodal alternatives.

Multimodal text, image, and metadata configurations are best within the traditional and CLIP families across granularities. With traditional features, the multimodal configuration improves over the best unimodal score by 0.08 macro-F1 in the binary and three-class settings, and by 0.04 in the five-class setting. With CLIP, multimodal gains range from 0.07 to 0.09 macro-F1. By contrast, SBERT and EfficientNet show no meaningful multimodal gain. The unimodal results show no stable text or image dominance, because their ordering changes across representation families and target granularities. The main result is therefore not a general ranking of text and image, but the advantage of combining text, image, and metadata in the traditional and CLIP families.

5.3 Fusion Mechanism and Label Granularity

The fusion comparison tests whether cross-attention improves over concatenation on a reduced subset of 25,000 training, 5,000 validation, and 5,000 test posts. All methods use the same subset and CLIP-based inputs; the compared mechanisms are concatenation, direct cross-attention, and fine-tuned cross-attention. In this setting, concatenation attains the highest macro-F1 across all three formulations, as reported in Table 4. The gaps over direct cross-attention are 0.02, 0.05, and 0.02 in the binary, three-class, and five-class settings, while those over fine-tuned cross-attention are 0.06, 0.07, and 0.05. Without repeated runs or uncertainty estimates, these differences are interpreted descriptively: the tested cross-attention variants show no advantage over concatenation in this subset, but the result does not support general conclusions about fusion mechanisms. The highest macro-F1 decreases from 0.68 to 0.50 and 0.33 across the three tasks. Since the number of classes and chance level also change, these scores are not directly normalized measures of task difficulty, but record lower observed performance in the finer-grained formulations under the same protocol.

The results provide the answers to the four evaluation questions introduced in Section 1. For *Q1*, multimodal configurations improve over text-only and image-only alternatives in the traditional and CLIP families, but provide no meaningful gain in the SBERT + EfficientNet family. For *Q2*, traditional and vision–language representations yield essentially aligned highest observed scores, while the semantic representation remains lower. The small differences between traditional and vision–language results are not interpreted as evidence of superiority without repeated runs or uncertainty estimates. The answer to *Q3* is restricted to the reduced-subset CLIP experiment: concatenation yields higher observed scores than the two tested cross-attention mechanisms, but the result remains descriptive. For *Q4*, observed macro-F1 is lower in the three-class and five-class formulations, although scores across different numbers of classes are not directly comparable. These findings motivate Section 6, which analyzes the views and feature groups supporting the selected configuration.

6 Explainability Analysis of Engagement Prediction

We select for explainability the XGBoost model with concatenated CLIP text, image, and metadata features. It has the highest observed macro-F1 in the three- and five-class settings and ties the traditional multimodal model in the binary setting; without repeated runs or uncertainty estimates, this does not establish superiority. The analysis combines global SHAP attributions by view and metadata group with local perturbation maps.

Global View Contributions. We assess how the three views contribute to the selected CLIP + metadata model. Table 5 reports view-level SHAP shares for test samples of $n = \{5{,}000, 10{,}000, 20{,}000, 50{,}000\}$, allowing stability across explanation sizes to be checked. The shares remain stable. Text contributes least, at about 16–17%. Image is largest in the binary and three-class settings, at about

Table 5: Global view contributions. Values are rounded normalized SHAP shares.

	Binary			**3-class**			**5-class**		
n	Text	Image	Meta	Text	Image	Meta	Text	Image	Meta
5,000	0.17	0.43	0.41	0.17	0.42	0.41	0.16	0.41	0.43
10,000	0.17	0.43	0.40	0.17	0.42	0.41	0.16	0.41	0.43
20,000	0.17	0.43	0.40	0.17	0.42	0.41	0.16	0.41	0.43
50,000	0.17	0.43	0.41	0.17	0.42	0.41	0.16	0.41	0.43

Table 6: Class-specific view attribution patterns.

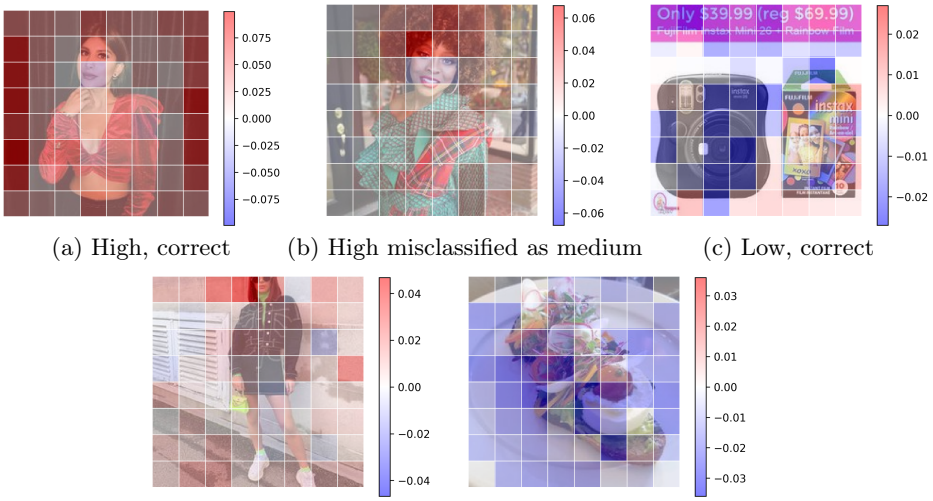
Setting	**Class pattern**	**Text**	**Image**	**Meta**
Binary	`low`/`high`	0.17	0.42–0.44	0.39–0.42
3-class	`low`/`high`	0.14–0.15	0.40–0.42	0.42
3-class	`medium`	0.30	0.44	0.26
5-class	`very_low`/`very_high`	0.13–0.15	0.38–0.39	0.47–0.48
5-class	`medium`	0.30	0.44	0.26

43% and 42%, while metadata is largest in the five-class setting ($\sim$ 43%). Image and metadata dominate, while caption text contributes less across settings.

Since individual CLIP text and image dimensions are not directly interpretable, metadata is considered separately. Dense dimensions are interpreted only after view-level aggregation, while feature-level analysis is limited to metadata; view shares thus represent model-level contributions. Across granularities, `posts` is the strongest metadata feature, with normalized shares of about 0.55, 0.28, and 0.31 in the binary, three-class, and five-class settings. The next variables are `followees` and `n_hashtags`. Together, `posts` and `followees` describe account structure, while `n_hashtags` describes caption structure. However, `posts` and `followees` are static snapshots, and the temporal split is not creator-disjoint. Their attributions may partly reflect stable creator signals or account states measured outside posting time. The result applies to observed creators and does not show generalization to unseen ones.

Class-Level Attribution Patterns. Class-specific SHAP aggregations show that the view-level pattern changes with the predicted class. Table 6 summarizes the numerical patterns reported for binary classes, three-class extremes, medium classes, and five-class extremes. In the binary setting, `low` and `high` engagement show similar shares. Image ranks first, metadata follows closely, and text remains smallest. Thus, the low-versus-high distinction is driven mainly by visual and structured information, not captions.

In the three-class setting, metadata is more relevant for the extremes. For `low` and `high`, metadata contributes about 0.42, image about 0.40–0.42, and text about 0.14–0.15. The `medium` class differs. Text rises to about 0.30, metadata decreases to about 0.26, and image remains largest, at about 0.44. This suggests



(a) High, correct (b) High misclassified as medium (c) Low, correct

(d) Medium, correct (e) Medium misclassified as low

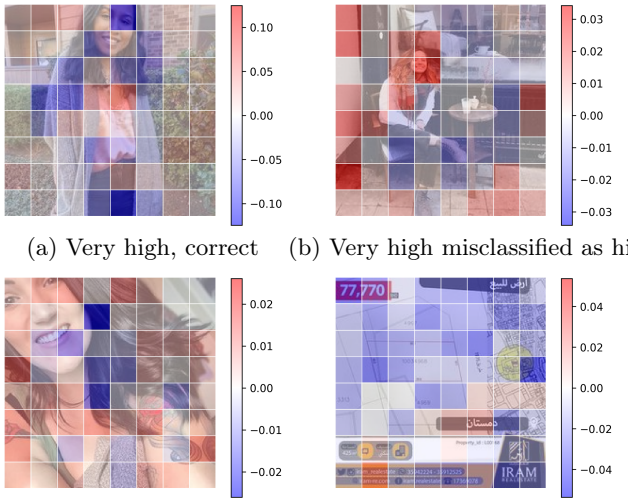
Fig. 2: Local image explanation examples for three-class predictions.

that medium-engagement posts are more heterogeneous and less anchored to account-level structure than the extremes.

The five-class setting reinforces this pattern. For `very_low` and `very_high`, metadata becomes the leading view, reaching about 0.47–0.48 of the attribution. Image remains important, while text provides the smallest share. The `low` and `high` classes follow the same ordering as in the three-class setting, while intermediate classes show a more mixed attribution profile. Thus, account-level structural features are especially relevant when the model distinguishes very low and very high engagement. Given the lower observed macro-F1 in the five-class setting, these attributions describe how the fitted model separates fine-grained labels, not stable estimates of engagement mechanisms. The stronger metadata contribution at the extremes may also partly reflect stable creator-specific information captured by the account-level variables.

Local Image Explanations. Global SHAP values summarize the model over posts, but not how individual predictions are formed. We therefore extend the global analysis with perturbation-based local explanations. MIXER supports perturbations of image regions and caption spans; here, image-region maps summarize local model sensitivity, while caption perturbations are used only as supporting qualitative evidence. The examples are qualitative diagnostics; the quantitative explanation evidence is provided by the global SHAP analysis.

For space reasons, we report local image maps for the three-class and five-class settings only. The selected examples include representative correct predictions and errors between adjacent classes. The binary setting follows the same broad pattern as the global analysis, with correct low-versus-high predictions



(a) Very high, correct (b) Very high misclassified as high

(c) Very low, correct (d) Very low misclassified as low

Fig. 3: Local image explanations for extreme classes in the five-class setting.

relying on visual regions together with metadata and caption perturbations producing smaller changes. However, it is less informative locally because it collapses neighbouring engagement levels into two broad classes. The three- and five-class settings better expose ambiguous cases, adjacent-label errors, and extreme engagement classes. For each selected post, the image is divided into a grid, one region is masked at a time, and the change in predicted-class probability is measured. Positive regions support the predicted class; negative regions inhibit it.

In the three-class setting, Figure 2 illustrates how local evidence changes between correct predictions and adjacent-class errors. Panels (a) and (c), correctly predicted as `high` and `low`, show more coherent local support for the predicted class. Panel (d), correctly predicted as `medium`, is less clear-cut, suggesting that medium engagement is harder to characterize locally. The errors in panels (b) and (e), from `high` to `medium` and from `medium` to `low`, show cases in which the available image evidence does not separate neighbouring classes strongly enough.

In the five-class setting, Figure 3 focuses on the extreme classes, where global SHAP assigns a larger role to metadata. Panels (a) and (c), correctly predicted as `very_high` and `very_low`, show that local image regions may support the prediction, but the decision can still depend strongly on metadata. Panels (b) and (d) show errors from `very_high` to `high` and from `very_low` to `low`. The displayed errors remain within adjacent classes rather than spanning opposite engagement levels.

7 Discussion and Limitations

Predictive results show clear multimodal gains in the traditional and CLIP families, while the semantic family remains lower. Traditional and vision–language configurations yield essentially aligned scores; the 0.01 differences in the three-class and five-class settings do not establish superiority without repeated runs or uncertainty estimates. These findings agree with prior Instagram research showing that image, text, and influencer-level information jointly support engagement prediction, with image and account features contributing more than text [3]. MIXER extends this evidence through a common comparison across three representation families and target formulations, followed by global and local post-hoc explanations. Its contribution lies mainly in the integrated protocol and systematic comparison, while the multimodal gains and lower text contribution confirm earlier findings. The highest macro-F1 values are 0.69, 0.52, and 0.34 across the three tasks. The lower values for more classes are descriptive because macro-F1 is not chance-normalized across class counts, and the evaluation lacks ordinal metrics.

The explainability analysis shows how the fitted model uses the available information. Text has the smallest attribution share, around 16–17%, while image and metadata account for most attributions. In the five-class setting, especially at the extremes, metadata becomes the strongest view, reaching about 48%. The leading metadata variables are `posts`, `followees`, and `n_hashtags`; the first two describe account structure and the last caption structure. Although likes, comments, and followers are excluded, account variables may capture account maturity, posting history, stable creator identity, or later account state.

The intended scenario is temporal prediction of later posts for observed creators. The split is chronological by post but not creator-disjoint, and account variables are static snapshots rather than post-aligned measurements. Their attributions are therefore setting-specific and do not show generalization to unseen creators or forecasting before publication.

The results represent predictive associations rather than causal effects. They show that `posts`, `followees`, and `n_hashtags` support prediction under the adopted data, target, and protocol, not that they cause engagement. Engagement is observed, whereas impressions, reach, ranking decisions, and recommendation mechanisms are not; the paper therefore addresses an observable user-interaction outcome rather than direct platform amplification.

The dataset lacks protected or group-level attributes, so the analysis cannot address demographic fairness. Follower normalization and training-only category balancing address audience-size and category imbalance only. A fairness analysis would require group data, exposure measures, and a dedicated protocol.

The limitations concern data, evaluation, and model formulation. The dataset contains English-captioned posts from 2017–2018, so the patterns may not describe current or multilingual Instagram dynamics. Follower count is the most recent available value and may underestimate historical engagement rates for accounts that gained followers. The cross-attention comparison uses a reduced subset and lacks repeated runs or confidence intervals, so its differences remain

empirical trends rather than statistically tested effects. Category-specific results and group-level fairness analyses are also absent. Finally, the ordered engagement classes are modeled and evaluated as nominal, without ordinal classifiers, losses, or metrics such as class-distance error or weighted agreement.

Future work should use newer, creator-disjoint datasets, time-aligned account features, ordinal models and metrics, category-level comparisons, and exposure or group-level information when available. These extensions would help separate post-level information from stable creator signals and assess whether predictive and attribution patterns persist across creators, periods, and engagement levels.

8 Conclusion

This paper presented MIXER, an integrated protocol for multimodal engagement prediction and post-hoc explanation of feature contributions in image-based social media posts. We evaluated it on Instagram data by comparing traditional, semantic, and vision–language representations across binary, three-class, and five-class tasks. The CLIP-based configuration selected for explainability combines text embeddings, image embeddings, and metadata through concatenation, attaining macro-F1 values of 0.69, 0.52, and 0.34. The results show that multimodal configurations improve over unimodal alternatives in the traditional and CLIP families. Observed macro-F1 is lower in the three-class and five-class tasks, but cross-granularity values are interpreted descriptively because class count and chance level change. The explainability analysis shows that text contributes least, while image and metadata account for most attributions. In the five-class setting and at the extremes, metadata becomes the strongest view, with `posts` and `followees` among the most influential variables. Because these variables are static account snapshots and the split is not creator-disjoint, their attribution may partly reflect stable creator-specific signals.

MIXER shows that predictive performance and feature attributions should be jointly analyzed: performance identifies effective models, while attributions clarify which views and feature groups support predictions. Future work should test newer, creator-disjoint datasets, use time-aligned account information, include ordinal evaluation, compare content categories, and connect engagement prediction with exposure or group-level information when available. These extensions would help separate post-level contributions from stable creator signals and assess whether the observed patterns generalize across creators, periods, and engagement levels.

Acknowledgments. This work was partially supported by the Italian Project Fondo Italiano per la Scienza FIS00001966 “MIMOSA”, and by the European Union under the scheme HORIZON-INFRA-2025-01-DEV-02, Grant Agreement No. 101292277, “SoBigData IP: SoBigData Implementation Phase”.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Carta, S., et al.: Popularity prediction of instagram posts. Information **11**(9) (2020). https://doi.org/10.3390/info11090453
2. Cruz, J.P.O., et al.: Characterizing fashion influencers' behavior on instagram. Social Network Analysis and Mining **14**(1), 147 (2024). https://doi.org/10.1007/s13278-024-01313-x
3. Hsiao, Y.H., Lin, Y.Y.: Decoding influencer marketing effectiveness on instagram: Insights from image, text, and influencer features. Journal of Retailing and Consumer Services **85**, 104285 (2025)
4. Kim, S., et al.: Multimodal post attentive profiling for influencer marketing. In: Proceedings of The Web Conference 2020 (WWW'20) (04 2020). https://doi.org/10.1145/3366423.3380052
5. Mazloom, M., et al.: Multimodal popularity prediction of brand-related social media posts. In: MM '16: Proceedings of the 24th ACM international conference on Multimedia. pp. 197–201. ACM (2016). https://doi.org/10.1145/2964284.2967210
6. Mazloom, M., et al.: Category Specific Post Popularity Prediction, pp. 594–607. Springer (01 2018). https://doi.org/10.1007/978-3-319-73603-7_48
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763 (2021)
8. Riis, C., et al.: On the limits to multi-modal popularity prediction on instagram: A new robust, efficient and explainable baseline. In: Rocha, A.P., Steels, L., van den Herik, H.J. (eds.) Proceedings of the 13th International Conference on Agents and Artificial Intelligence, ICAART 2021, Volume 2, Online Streaming, February 4-6, 2021. pp. 1200–1209. SCITEPRESS (2021). https://doi.org/10.5220/0010377112001209
9. Rivadeneira, L., et al.: Evidential reasoning approach for predicting popularity of instagram posts. IEEE Access **12**, 182603–182617 (2024). https://doi.org/10.1109/ACCESS.2024.3510637
10. Tricomi, P.P., et al.: Follow us and become famous! insights and guidelines from instagram engagement mechanisms (2023), `https://arxiv.org/abs/2301.06815`
11. Viola, M., et al.: Instagram images and videos popularity prediction: a deep learning-based approach. In: Proceedings of the 1st Italian Workshop on Artificial Intelligence and Applications for Business and Industries (AIABI 2021), co-located with AI*IA 2021. vol. 3102 (2021), `https://ceur-ws.org/Vol-3102/paper2.pdf`
12. Zohourian, A., et al.: Popularity prediction of images and videos on instagram. In: 2018 4th International Conference on Web Research (ICWR). pp. 111–117 (2018). https://doi.org/10.1109/ICWR.2018.8387246

A Experimental Details

Preprocessing. Posts are restricted to 2017–2018 observations with English captions, at least one valid image, and a valid metadata–image mapping. Captions are cleaned by removing hyperlinks, email addresses, and user mentions, while hashtags are reduced to their textual component. For TF–IDF, stopwords are also removed and the vocabulary is fitted on training data only. Images are converted to RGB and resized to 224×224. For posts with multiple images, image-level representations are mean-pooled into one post-level vector. Categorical metadata variables are encoded as numerical indicators and numerical variables are standardized. All data-dependent steps are fitted on training data and applied unchanged to validation and test data.

Split and balancing protocol. A chronological split avoids mixing past and future posts: training covers January 2017–October 2018, validation November 2018, and test December 2018. Validation is used for model selection and hyperparameter tuning, while test data are held out for final reporting. Category balancing is applied only to training. The smallest creator category, *pet*, contains 544 influencers and defines the reference size. For each other category, 544 influencers are selected through a reproducible pseudo-random procedure based on sorted username hashes, retaining all their posts. Validation and test remain unchanged, preserving their natural distributions.

B Feature Details

Metadata features. Table 7 reports the metadata feature groups used in the experiments. The predictive feature set includes temporal, caption-structure, image-geometry, creator-category, and account-level structural variables. Target-related variables are excluded to avoid direct leakage from the engagement-rate definition, i.e., `like_count`, `comment_count`, `followers`, and `engagement_rate`.

Text and image representations. The traditional family represents captions with TF–IDF unigram vectors, using `min_df`$= 20$, `max_df`$= 0.7$, a maximum of 50,000 features, and category-specific top-word indicators. Its image view is represented

Table 7: Metadata feature groups used in the experiments.

Group	**Variables**
Temporal	`year`, `month`, `dow`, `hour_utc`
Caption structure	`has_caption`, `caption_len_char`, `n_hashtags`, `n_mentions`, `n_urls`, `n_emojis`
Image geometry	`width`, `height`, `aspect_ratio`, `area`, `orientation`
Creator category	`category`
Account-level structural	`followees`, `posts`
Excluded target-related	`like_count`, `comment_count`, `followers`, `engagement_rate`

Table 8: Classifier-level macro-F1 by modality and representation; bold marks row maxima.

Family	Configuration	Linear SVM	NB	RF	XGBoost
Traditional	Text, binary	**0.61**	0.54	0.54	0.55
Traditional	Text, 3-class	**0.43**	0.36	0.29	0.35
Traditional	Text, 5-class	**0.29**	0.22	0.17	0.22
Traditional	Image, binary	0.55	0.49	0.55	**0.57**
Traditional	Image, 3-class	0.38	0.30	0.38	**0.39**
Traditional	Image, 5-class	0.23	0.17	0.23	**0.25**
Traditional	Multimodal, binary	0.68	0.58	0.67	**0.69**
Traditional	Multimodal, 3-class	0.47	0.40	0.49	**0.51**
Traditional	Multimodal, 5-class	0.23	0.23	0.32	**0.33**
SBERT + EfficientNet	Text, binary	0.57	0.55	0.57	**0.58**
SBERT + EfficientNet	Text, 3-class	0.38	0.37	0.39	**0.40**
SBERT + EfficientNet	Text, 5-class	0.24	0.22	0.23	**0.25**
SBERT + EfficientNet	Image, binary	**0.59**	0.57	0.58	**0.59**
SBERT + EfficientNet	Image, 3-class	**0.41**	0.37	**0.41**	**0.41**
SBERT + EfficientNet	Image, 5-class	0.25	0.21	0.25	**0.26**
SBERT + EfficientNet	Multimodal, binary	**0.60**	0.58	0.59	**0.60**
SBERT + EfficientNet	Multimodal, 3-class	0.41	0.38	0.41	**0.42**
SBERT + EfficientNet	Multimodal, 5-class	0.25	0.21	0.23	**0.26**
CLIP	Text, binary	**0.58**	0.55	0.57	**0.58**
CLIP	Text, 3-class	**0.41**	0.37	0.38	**0.41**
CLIP	Text, 5-class	**0.25**	0.21	0.22	**0.25**
CLIP	Image, binary	**0.62**	0.60	0.61	**0.62**
CLIP	Image, 3-class	0.42	0.40	0.43	**0.44**
CLIP	Image, 5-class	0.16	0.08	**0.19**	0.17
CLIP	Multimodal, binary	0.68	0.63	0.67	**0.69**
CLIP	Multimodal, 3-class	0.45	0.44	0.50	**0.52**
CLIP	Multimodal, 5-class	0.30	0.26	0.32	**0.34**

by 149 handcrafted descriptors, including colorfulness, Hue Saturation Lightness statistics and histograms, grayscale histograms, entropy, Laplacian variance, edge density, Local Binary Patterns, and Gray-Level Co-Occurrence Matrix features. The semantic family uses 384-dimensional SBERT caption embeddings and 1280-dimensional EfficientNet-B0 image embeddings. The vision–language family uses CLIP text and image embeddings, both with 512 dimensions.

C Model Details

The experiments compare four classifier families: Linear SVM as a linear margin-based classifier, Gaussian Naive Bayes as a probabilistic baseline, Random Forest as a bagging-based ensemble, and XGBoost as a gradient-boosted tree model. Only the hyperparameters required for model selection are tuned, keeping the validation protocol computationally bounded and consistent across representation families. Table 8 reports classifier-level test macro-F1 for the configurations used in Table 3. Its multimodal rows support the comparison across representation families, while its text-only, image-only, and multimodal rows together support the within-family comparison across modality settings.