FOCAL Ad: Multimodal Contrastive Learning for Robust Advertising Conversion Prediction in Factorized Orthogonal Latent Space

N V Navaneeth Rajesh, Kshitij Vaidya, Raunak Mukherjee
22b1215@iitb.ac.in, 22b1829@iitb.ac.in, 22b3955@iitb.ac.in

Abstract

In digital advertising, predicting conversion rates relies heavily on understanding the synergy between visual creatives (images) and textual copy (captions). However, these modalities often contain modality-specific noise, such as visual clutter or linguistic stylistic flourishes, that confuse predictive models. Additionally, visual creatives and textual copies transmit different kinds of information to the consumer. Standard multimodal fusion techniques often fail to disentangle the core semantic message common to both modalities from the information private to each modality. This private information may either be modality-specific or truly informative features. In this paper, we propose **FOCAL Ad** an adaptation of the **FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space** [4] framework to the novel domain of advertising. By enforcing a strict factorization of the latent space into *Shared* (Semantic) and *Private* (Style) subspaces via orthogonality constraints, we extract robust features that represent the coherent ad message as well as modality-private information transmitted uniquely through specific modalities. We demonstrate this method on the Flickr8k dataset [3], adapted as a proxy for ad creatives, and show that training on factorized shared and private embeddings yields robust performance on conversion prediction tasks compared to a purely shared baseline.

1 Introduction

The effectiveness of a digital advertisement is rarely defined by a single modality. It is the alignment between the *visual hook* (the image) and the *value proposition* (the text) that drives user conversion. In machine learning, this is a classic multimodal fusion problem.

However, real-world data is noisy. An image may contain lighting variations or background clutter (visual noise), while ad copy may use varying degrees of verbosity or slang (textual noise). Standard contrastive learning methods, such as CLIP [5], attempt to pull the entire embedding of an image and its text closer together. This inevitably forces the model to encode noise into the shared representation, degrading performance on downstream tasks like Click-Through Rate (CTR) prediction. Furthermore, each modality equips the advertiser to transmit information in unique fashions that are particular to the modality. For example, consider food advertising, food adverts often feature appealing photographs that can uniquely provide a visual representation as compared to a simply text / slogan based campaign. However, catchy slogans may be the differentiating factor between two food companies that sell similar products. Hence, each modality may contain useful private features that drive CTR as well.

To address this, we propose **Focal Ad**, an adaptation and application of **FOCAL: Contrastive Learning for Multimodal Time-Series Sensing Signals in Factorized Orthogonal Latent Space** [4]. FOCAL differs from standard fusion by explicitly decomposing the feature space. It assumes that for any ad, there exists:

- A **Shared Component** (S): The core concept (e.g., "Running Shoes") present in both image and text.
- A **Private Component** (P): Modality-specific attributes (e.g., "Sunset lighting" in the image, or "Exclamation marks" in the text).

By enforcing mathematical *orthogonality* between S and P, we ensure that the S embedding used for conversion prediction is clean, semantically dense, and robust to stylistic noise. We run an experiment on the Flickr8k dataset [3] and demonstrate that using both Shared and Private Components boosts model performance and provide visualizations of the latent space.

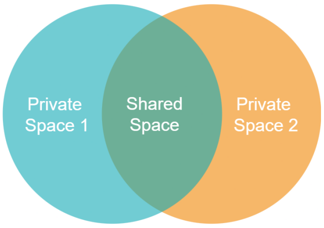

Figure 1: Diagram illustrating the Shared and Private Factorized Latent Spaces for Hypothetical Modalities 1 and 2.

For example, consider two hypothetical modalities 1 and 2. We would have a shared component embeddings in S and private component embeddings corresponding to modality 1 and modality 2 in $P1$ and $P2$. See figure 1 for a visual illustration.

2 Related Works

Contrastive learning has emerged as a powerful framework for self-supervised representation learning. Foundational visual contrastive methods such as SimCLR [1] and MoCo [2] showed that instance-level alignment and carefully selected negatives can yield high-quality semantic embeddings without labels. Multimodal extensions, most notably CLIP [5], demonstrated that contrastive learning between images and text can produce robust *shared semantic spaces* at scale. In parallel, research in time-series contrastive learning, including TS2Vec [6] highlighted the importance of temporal adjacency and structured augmentations when learning from sequential sensor data.

FOCAL builds on these threads by proposing a **factorized latent space** with shared and private subspaces, using contrastive objectives and orthogonality constraints to separate modality-invariant semantics from modality-specific information. This design is tailored for multimodal time-series sensing, where augmentations preserve semantic content and temporal structure provides natural positive pairs. Our work differs in that advertising data consists of **static** image–caption pairs: textual augmentations often alter meaning, many image augmentations disrupt critical visual content, and *no temporal structure exists.* Consequently, we retain only the components of FOCAL compatible with static multimodal data—shared-space alignment and orthogonal shared/private disentanglement—while omitting augmentation-dependent and sequence-dependent objectives. This yields a simplified yet principled variant suitable for robust multimodal advertising prediction.

3 FOCAL: Recap of Full Methodology

FOCAL learns a factorized latent space with a *shared* subspace and a *private* subspace for each modality. Each sample i in modality j yields two outputs: h_{ij}^{sh} (shared) and h_{ij}^{pr} (private).

To keep equations compact in two-column format, we define a short scoring function:

$$\phi(u, v) = \exp(\langle u, v\rangle/\tau),$$

and a normalized contrastive loss for a positive pair (u, v) with negative set $\mathcal{Z}$:

$$\mathcal{C}(u, v; \mathcal{Z}) = -\log \frac{\phi(u, v)}{\sum_{z\in\mathcal{Z}} \phi(u, z)}.$$

1. Shared-space contrastive loss $\mathcal{L}_{\mathbf{shared}}$. Positives are shared embeddings of the same sample across modalities:

$$\mathcal{L}_{\text{shared}} = \sum_i \sum_{j\neq j'} \mathcal{C}\big(h_{ij}^{\text{sh}}, h_{ij'}^{\text{sh}}, \mathcal{N}_{ij}^{\text{sh}}\big),$$

where $\mathcal{N}_{ij}^{\text{sh}}$ is the set of shared negatives within the batch.

2. Private-space contrastive loss $\mathcal{L}_{\mathbf{private}}$. For each modality, the paired augmentation $(h_{ij}^{\text{pr}}, \tilde{h}_{ij}^{\text{pr}})$ is a positive pair:

$$\mathcal{L}_{\text{private}} = \sum_i \sum_j \mathcal{C}\Big(h_{ij}^{\text{pr}}, \tilde{h}_{ij}^{\text{pr}}, \mathcal{N}_{ij}^{\text{pr}}\Big),$$

with $\mathcal{N}_{ij}^{\text{pr}}$ containing private negatives from the same modality.

3. Orthogonality constraint $\mathcal{L}_{\mathbf{ortho}}$. We enforce decorrelation between shared and private spaces and across-modality private spaces. Let

$$\psi(u, v) = \langle u, v\rangle,$$

then

$$\mathcal{L}_{\text{ortho}} = \sum_i \sum_j \psi(h_{ij}^{\text{sh}}, h_{ij}^{\text{pr}}) + \sum_i \sum_{j\neq j'} \psi(h_{ij}^{\text{pr}}, h_{ij'}^{\text{pr}}).$$

4. Temporal structural loss $\mathcal{L}_{\mathbf{temporal}}$. Let $\bar{D}_{ss}$ be the mean intra-sequence distance and $\bar{D}_{ss'}$ the mean inter-sequence distance. With margin μ:

$$\mathcal{L}_{\text{temporal}} = \sum_s \sum_{s'\neq s} \max(\bar{D}_{ss} - \bar{D}_{ss'} + \mu,\ 0).$$

Overall pretraining objective.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{shared}} + \lambda_p \mathcal{L}_{\text{private}} + \lambda_o \mathcal{L}_{\text{ortho}} + \lambda_t \mathcal{L}_{\text{temporal}}.$$

FOCAL uses MLP projectors for shared/private heads, modality-appropriate augmentations, and batches composed of multiple short sequences to compute sequence-level structure.

4 FOCAL Ad: Our Methodology

This section describes the design deviations from the full FOCAL framework in order to adapt the method to the realm of multimodal advertising. Details in this section reflect the executed Python implementation.

4.1 Problem Formulation

Let $\mathcal{D} = \{(x_i^A, x_i^B)\}_{i=1}^N$ denote image–caption pairs from the Flickr8k dataset. For each pair we extract modality-specific high-level features and learn a factorized encoder producing (i) a *shared* embedding capturing semantic information common to both modalities, and (ii) a *private* embedding capturing modality-specific signals. These embeddings are trained to be useful for a synthetic downstream binary task $Y \in \{0, 1\}$ representing simulated conversion likelihood.

4.2 Feature Extraction

In our implementation, feature extraction is *not* performed end-to-end during training. Instead, we precompute high-level features once and cache them to disk. This design significantly reduces training time.

- **Image features:** Extracted using a ResNet-18 model pre-trained on ImageNet, with the classification head removed, yielding $h_A \in \mathbb{R}^{512}$.
- **Text features:** Extracted using BERT-base-uncased, where the `[CLS]` token embedding yields $h_B \in \mathbb{R}^{768}$.

These cached feature vectors are fed directly into the FOCAL encoder.

4.3 Implemented FOCAL Encoder

For each modality $m \in \{A, B\}$, the encoder applies a feed-forward MLP followed by two projection heads:

$$h'_m = \text{MLP}(h_m), \tag{1}$$

$$s_m = \text{Head}_{\text{shared}}(h'_m), \qquad p_m = \text{Head}_{\text{private}}(h'_m), \tag{2}$$

with $s_m, p_m \in \mathbb{R}^{128}$ in our implementation.

Unlike the full FOCAL paper, which includes contrastive objectives for both subspaces and temporal constraints, our implementation simplifies and adapts the objective to only use the losses described below.

4.4 Objective Functions

4.4.1 1. Shared Similarity Loss

We enforce semantic consistency by minimizing cosine distance between shared embeddings:

$$\mathcal{L}_{\text{sim}} = 1 - \frac{s_A \cdot s_B}{\|s_A\| \, \|s_B\|}. \tag{3}$$

This encourages shared embeddings of image–caption pairs to align.

4.4.2 2. Orthogonality Loss

To decouple semantic and modality-specific components, we apply a soft orthogonality constraint:

$$\mathcal{L}_{\text{ortho}} = \frac{1}{2}\left(|s_A^\top p_A| + |s_B^\top p_B|\right). \tag{4}$$

This term ensures s_m does not encode modality-specific information captured by p_m.

4.4.3 3. Total Training Loss

Our implemented total loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sim}} + \lambda\, \mathcal{L}_{\text{ortho}}, \tag{5}$$

with $\lambda = 0.05$.

4.5 Deviations from the FOCAL Paper

We highlight our deviations from the original FOCAL methodology:

- **No private NT-Xent contrastive loss.** The full FOCAL design requires augmentations of each input and a private-space contrastive loss. However, in our case, simple text augmentations (synonym replacement, masking) change semantics, not style. Because FOCAL assumes augmentations that preserve underlying meaning but alter modality-specific components, most NLP augmentations violate this assumption.
- **No temporal or sequence-level constraint.** Flickr8k does not contain sequences, trajectories, or temporal evolution. Applying temporal consistency loss to unordered independent samples is mathematically unjustified.
- **No cross-augmentation pairing.** For reasons similar to the omitting the private NT-Xent loss, generating "two valid views" is not well-defined for text captions. Most text perturbations alter semantic meaning or grammar, violating FOCAL's assumption that augmentations preserve semantics.

These omissions do not invalidate the factorized-latent-space idea, but they should be clearly stated since our implementation is a modified simpler *FOCAL variant* more apt for the realm of advertising prediction. We preserve the key components relevant for multimodal alignment: shared-space similarity and shared/private disentanglement.

4.6 Downstream Task and Synthetic Labeling

A synthetic "conversion" label is generated using a hidden linear projection over the concatenated features:

$$\text{logits} = W[h_A; h_B] + b + \epsilon,$$

and thresholding at the median to ensure class balance. This makes the downstream task dependent on both modalities while inserting Gaussian noise for robustness.

4.7 Training Details

We train for 30 epochs using Adam (LR $= 10^{-4}$) with batch size 128. Only the encoder MLPs are trained; feature extractors remain frozen because features were precomputed.

5 t-SNE Embedding Analysis

We use t-SNE to visualize various learned representations. t-SNE is applied directly to the learned embeddings without PCA, reflecting the implementation.

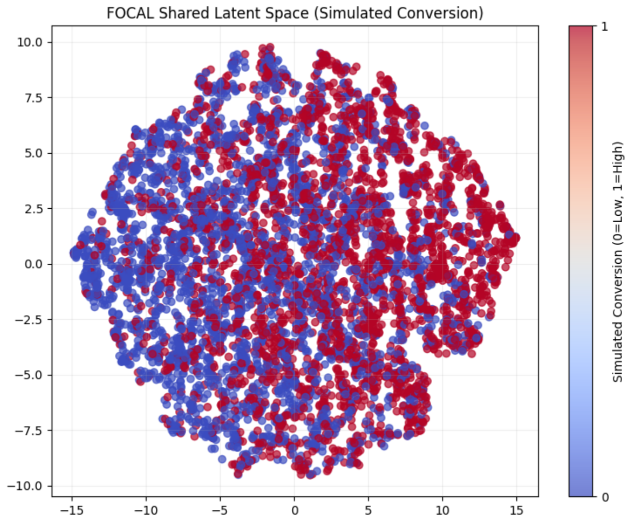

Figure 2: t-SNE projection of the FOCAL Shared Latent Space on the Test set. Points are colored by the Conversion Label (Red=High, Blue=Low). The clear separation indicates that the Shared Space successfully captures the features driving conversion.

5.1 t-SNE Procedure

For any embedding matrix $Z \in \mathbb{R}^{N \times d}$, t-SNE maps Z to a 2-D space by constructing a low-dimensional distribution that preserves local neighbor structure. Distances are modeled using conditional probabilities:

$$p_{j|i} = \frac{\exp(-\|z_i - z_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|z_i - z_k\|^2/2\sigma_i^2)},$$

while the 2-D embeddings use a heavy-tailed Student-t distribution:

$$q_{j|i} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|y_i - y_k\|^2)^{-1}}.$$

The embedding minimizes the KL divergence:

$$\mathrm{KL}(P\|Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

This emphasizes preserving local neighborhoods—ideal for diagnosing structure in learned latent spaces.

6 Experimental Setup

6.1 Dataset: Flickr8k as Ad Proxy

Due to the proprietary nature of commercial Click-Through Rate (CTR) datasets containing raw images, we utilize the **Flickr8k** dataset [3] as a proxy.

- **Data Structure:** 8,000 images, each paired with 5 captions.
- **Ad Adaptation:** We treat the image as the "Ad Creative" and the captions as variations of "Ad Copy."

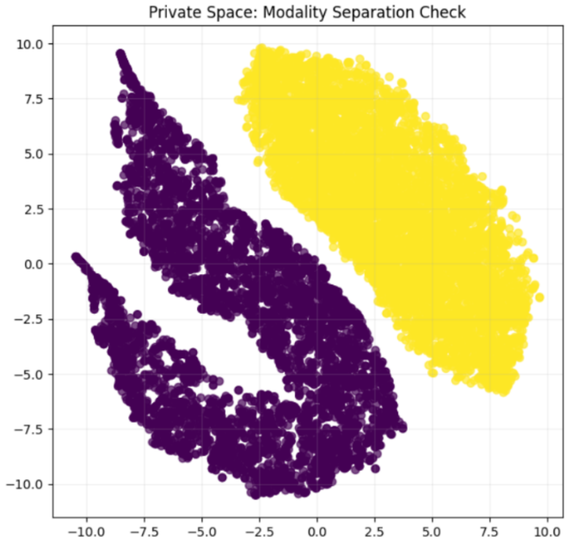

Figure 3: t-SNE projection of FOCAL Private Space Modality features on the Test set. Points are colored by modality (Yellow = Text Private, Purple = Image Private The clear separation successfully demonstrates that the private space encodes information unique to its modality.

- **Preprocessing:** Images are resized to 224 × 224. Captions are tokenized with a max length of 50.

6.2 Simulated Conversion Target

To rigorously evaluate the model's ability to fuse modalities for prediction, we generated a synthetic "Conversion" label (Y). The ground truth Y is generated via a hidden linear mapping of the concatenated raw features (h_A, h_B) with added Gaussian noise ϵ:

$$\text{Logits} = W \cdot [h_A; h_B] + b + \epsilon \tag{6}$$

$$Y = \mathbb{I}(\text{Logits} > \text{Median}(\text{Logits})) \tag{7}$$

This creates a balanced binary classification task where the signal exists in the fusion of the modalities, but is obfuscated by noise.

6.3 Training Details

We trained the FOCAL encoders using the **Adam** optimizer with a learning rate of $1e-4$. The batch size was set to 128, and training proceeded for 30 epochs. The latent dimension d for both shared and private spaces was set to 128.

7 Results and Discussion

7.1 Quantitative Analysis

We evaluated the utility of the learned embeddings on the downstream Conversion Prediction task using Logistic Regression. We compared the performance of using only the **Shared** embeddings (S_A, S_B) against using the **Full** embeddings (S_A, P_A, S_B, P_B).

As shown in Table 1, the **Full (Shared + Private)** model achieves superior performance as compared to the **FOCAL Shared Only** model. This confirms that the FOCAL objective successfully distilled the relevant predictive signal into the Shared space, while the Private space likely contained additional features unique to a modality. The boost in accuracy is not very large however, as we expect the Text Modality and the Image Modality to contain semantically similar information in a well-designed advertising campaign.

Table 1: Downstream Classification Accuracy

Embedding Source	**Latent Dim**	**Accuracy**
Full (Shared + Private)	**512**	**67.07%**
FOCAL Shared Only	256	66.28%
Chance Baseline	-	50.00%

7.2 t-SNE embedding pipeline and expected interpretations

All t-SNE visualizations were produced after a small preprocessing chain: (i) feature standardization; (ii) dimensionality reduction with PCA to $d_{\mathrm{PCA}} = 50$ (or $d_{\mathrm{PCA}} = \min(50, d)$ when $d < 50$); and (iii) 2D embedding with t-SNE (perplexity ≈ 30, $n\_iter = 1000$, init='pca'). PCA prior to t-SNE reduces noise and accelerates convergence, producing more stable layouts across runs.

We generate two diagnostic plots; the expected structure and interpretation are:

These visualizations collectively diagnose (i) semantic concentration in the shared space and (ii) modality-specificity in the private space.

Plot 1: FOCAL Shared Latent Space (per-sample concatenation $[S_A; S_B]$).

Expectation: If the shared heads learned modality-invariant semantics relevant to the downstream task, points should cluster by the target label Y (high vs low simulated conversion). Strong class separation indicates that the shared embedding encodes task-relevant information; heavy overlap indicates insufficient shared alignment or that private/style factors dominate.

Observation: The separation of the two clusters indicates that the shared heads learned modality-invariant semantics relevant to the downstream task. Although we note that the separation is not particularly large.

Plot 2: Private Space Modality Separation (stacked P_A then P_B, color = modality).

Expectation: Private vectors from image and text modalities should occupy distinct regions (two-mode separation) since each private head should encode modality-specific factors. If these overlap heavily, the private heads are not learning modality-exclusive signals and orthogonality or private contrastive objectives may be required.

Observation: The clear separation in figure 3 successfully demonstrates that the private space encodes information unique to its modality.

7.3 Latent Space Visualization

We visualized the learned Shared Latent Space using t-SNE (Figure 2).

The visualization demonstrates distinct clustering between the "High Conversion" and "Low Conversion" classes. Since the t-SNE was performed solely on the shared semantic embeddings, this implies that the "conversion" signal is semantically grounded—meaning successful ads in our simulation share common semantic traits (e.g., high-quality descriptions matching image content) rather than just stylistic quirks.

8 Conclusion

In this work, we presented a novel application of the FOCAL framework to multimodal advertising. By factorizing the latent space into orthogonal Shared and Private components, we successfully disentangled the core semantic message of an ad from its modality-specific noise. Our experiments on the Flickr8k dataset demonstrate that this factorization yields compact, robust embeddings that perform well on downstream conversion prediction tasks. Future work will involve applying this method to real-world Click-Through logs and incorporating user-history as a third temporal modality.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[3] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research (JAIR)*, 47:853–899, 2013.

[4] Shengzhong Liu, Tomoyoshi Kimura, Dongxin Liu, Ruijie Wang, Jinyang Li, Suhas Diggavi, Mani Srivastava, and Tarek Abdelzaher. Focal: Contrastive learning for multimodal time-series sensing signals in factorized orthogonal latent space, 2023.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[6] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series, 2022.