# FACTORIZED MULTIMODAL FUSION FOR DATA-CONSTRAINED REGIMES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal classification often requires substantial paired data to learn how to combine heterogeneous representations from different modalities. We consider a streamlined regime in which modality-specific encoders are frozen and adaptation operates directly within their pretrained representation spaces. Our method, **Fa**ctorized **M**odality **E**vidence (FAME), targets multimodal fusion in data-constrained settings through a factorized multimodality approach. We model class structure independently within each modality and combine the resulting evidence in the shared label space. Each modality contributes a Gaussian discriminant score, while a single scalar coefficient calibrates the modality evidence and class prior. This avoids representation-level alignment, encoder updates, and high-dimensional learned fusion, while naturally supporting missing modalities by omitting unavailable evidence terms. Across six datasets and eleven classification tasks, FAME outperforms the strongest competing baseline in 36 of 42 few-shot dataset–shot settings. Under test-time modality missingness, our method consistently outperforms standard imputation and modality-dropout baselines. When few-shot adaptation data contain missing modalities, FAME achieves the strongest performance across all 18 settings. These results show that lightweight factorized adaptation can effectively exploit complementary information from independently pretrained modalities without learning a high-capacity multimodal fusion model. Our code is available at this anonymized repository.

## 1 INTRODUCTION

Multimodal classification requires combining heterogeneous sources of information (Baltrušaitis et al., 2018; Lahat et al., 2015). Different modalities capture different aspects of an example, exhibit different statistical structure, and may contribute unequally to the same prediction task (Wang et al., 2020a;b). This heterogeneity also appears in modern pretrained encoders: strong image and text models are often trained independently with different objectives, producing representation spaces that can differ in dimension, geometry, and downstream predictive strength (Liang et al., 2022; Huo et al., 2024). These pretrained representations provide a strong starting point for multimodal learning, but adapting them to a downstream task can still require substantial labeled supervision (Cai et al., 2026).

This challenge is especially pronounced in data-constrained settings. Labeled multimodal examples can be expensive to obtain (Radford et al., 2021; Zhou et al., 2025), particularly when annotation requires domain expertise or coordinated collection across modalities, and complete modality information may not be available for every example (Wu et al., 2024). Few-shot supervision and missing modalities therefore arise naturally in practical multimodal settings, rather than only as adversarial stress tests (Ravi & Larochelle, 2017; Zhang et al., 2022). Under these constraints, updating large pretrained encoders or fitting a high-dimensional fusion model can be difficult because limited supervision must determine both the downstream class structure and how evidence from different modalities should be combined.

We ask whether multimodal adaptation can instead remain lightweight when each frozen modality already contains useful information about the target. Our approach models class structure independently within each pretrained representation space and combines the resulting evidence only in the shared label space. This keeps high-dimensional estimation local to each modality, while cross-



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

modal combination is reduced to a single scalar calibration coefficient. The resulting classifier can combine complementary evidence from independently pretrained encoders without representation-level alignment or encoder updates.

We develop FAME (**Fa**ctorized **M**odality **E**vidence) from this perspective. Under a conditional-independence assumption on the modality representations given the class, the joint posterior factorizes into a class prior and modality-specific likelihood terms (Kittler et al., 1998). We estimate each term with Gaussian discriminant analysis (GDA) (Bishop, 2006; Wang et al., 2024) over the frozen representations, yielding a closed-form class score within each encoder space. These modality-specific predictions are then combined with a prior-corrected product rule, with a single scalar controlling the combined modality evidence relative to the empirical class prior. Adaptation therefore consists of estimating class statistics independently within each frozen representation space together with one low-dimensional calibration parameter (Figure 1).

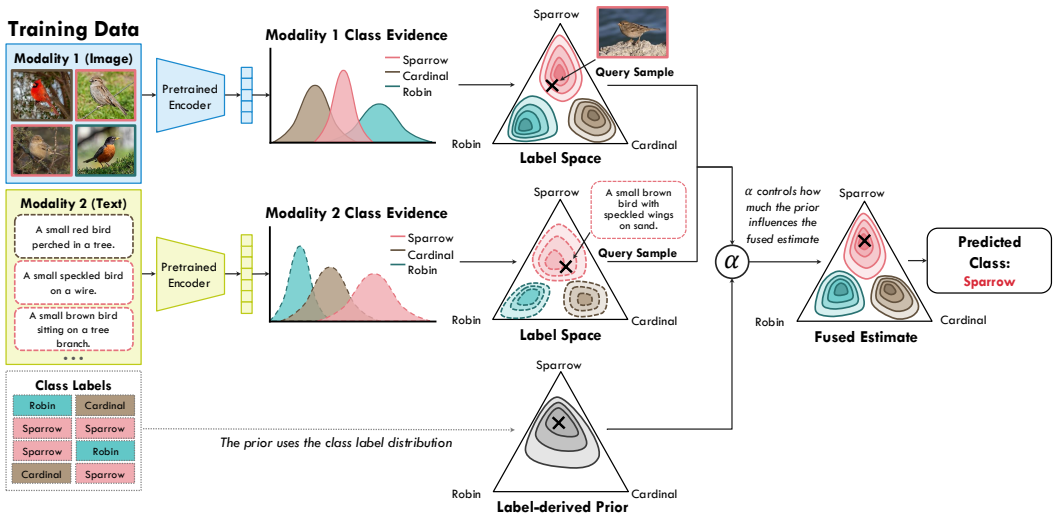



Figure 1: **Method overview.** FAME estimates modality-specific class evidence from frozen pretrained representations and combines it in the shared label space with an empirical class prior. The fused posterior is then used for classification.

This factorized structure is particularly useful when multimodal supervision is limited or incomplete. In few-shot settings, FAME avoids fitting a high-dimensional fusion model from a small labeled support set. When the adaptation data are incomplete, each modality-specific model can use every labeled example for which that modality is observed, rather than restricting estimation to fully paired examples. At inference time, an unavailable modality is handled by omitting its evidence term, allowing the same classifier to operate across different modality-availability patterns without feature imputation or retraining.

We evaluate FAME on six multimodal datasets spanning eleven classification tasks and consider several data-constrained regimes. In few-shot classification, FAME outperforms finetuning and other training-based baselines in $36/42$ settings. Under test-time modality missingness, it consistently improves over training-free alternatives and outperforms early-fusion approaches with standard imputation strategies by nearly $6\%$ points. When the few-shot adaptation set itself contains missing modalities, FAME outperforms training-based approaches in 18/18 evaluated settings, with improvements ranging up to $21.8\%$ points. Together, these results show that factorizing adaptation across pretrained modality spaces can be effective when labeled multimodal data are scarce or incomplete.

Our contributions are threefold:

- We formulate multimodal classification with frozen pretrained encoders as a lightweight adaptation problem, estimating class structure separately within each modality and combining evidence only in the shared label space.
- We introduce FAME, a gradient-free multimodal classifier that fits modality-specific Gaussian discriminant models and combines their predictions with a prior-corrected product rule

and a single scalar calibration coefficient. The method requires neither representation-level alignment nor encoder updates and naturally accommodates missing modalities.
- We evaluate FAME across six datasets and eleven tasks under few-shot supervision, test-time modality missingness, and incomplete few-shot adaptation data, where it performs strongly relative to both training-free and gradient-based baselines.

## 2 RELATED WORK

**Multimodal Learning and Fusion.** Multimodal learning studies how heterogeneous sources of information can be represented and combined for downstream prediction (Baltrušaitis et al., 2018; Lahat et al., 2015; Ngiam et al., 2011). Common approaches include early fusion of modality representations, late fusion of modality-specific predictions, and learned intermediate fusion. Representative methods include gated multimodal units (Arevalo et al., 2017), tensor and low-rank fusion (Zadeh et al., 2017; Liu et al., 2018), and cross-modal attention (Tsai et al., 2019). Recent work has also highlighted challenges arising from differences in modality informativeness, optimization dynamics, and generalization (Wang et al., 2020b; Peng et al., 2022; Fan et al., 2023).

**Efficient Adaptation of Vision-Language Models.** A large body of work studies lightweight adaptation of pretrained vision-language models, including prompt learning (Zhou et al., 2022b;a), adapters (Gao et al., 2024), and training-free methods (Zhang et al., 2021; Guo et al., 2023; Udandarao et al., 2023; Wang et al., 2024). Although these methods build on multimodal pretraining, they primarily target downstream image classification, where text represents class names or prompts rather than true inputs from multiple modalities. Moreover, these methods target CLIP embeddings (Radford et al., 2021) that already have a shared cross-modal embedding space. Related work such as ASIF (Norelli et al., 2023) connects independently pretrained image and text encoders without gradient-based training.

**Learning with Missing Modalities.** Robust multimodal prediction under incomplete observations is a well-motivated problem (Wu et al., 2024). Existing approaches typically account for missingness during training by reconstructing missing inputs or adding to the model's network. Popular approaches include modality dropout (Neverova et al., 2015), representation imputation or reconstruction (Wang et al., 2023; Zhang et al., 2022), modality-aware distillation (Wei et al., 2023), and meta-learning approaches (Ma et al., 2021).

**Discriminant Analysis and Classifier Combination.** Gaussian discriminant analysis (GDA) (Bishop, 2006; Fisher, 1936; Hastie & Tibshirani, 1996) models class-conditional feature distributions. In high-dimensional, low-sample regimes, covariance estimates are commonly stabilized through shrinkage or regularization (Friedman, 1989; Ledoit & Wolf, 2004; Chen et al., 2010). Separately, classical classifier-combination methods aggregate predictions or likelihoods from multiple representations using probabilistically motivated rules such as sums and products (Kittler et al., 1998).

## 3 METHOD

We consider classification into $C$ classes using $M$ modalities, indexed by $\mathcal{M} = \{1, \dots, M\}$. For each modality $m \in \mathcal{M}$, a pretrained encoder $f_m$ maps the input $x^{(m)}$ to a representation $z^{(m)} = f_m(x^{(m)}) \in \mathbb{R}^{d_m}$, where $d_m$ is the encoder's output dimension. We are given a support set of $N$ labeled examples, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, with $y_i \in \{1, \dots, C\}$. Here, $x_i$ denotes the multimodal input for example $i$, and $x_i^{(m)}$ denotes its observation from modality $m$, when available. Each example may contain all or only some of the modalities.

FAME uses this support set to fit a Gaussian classifier for each modality and then combines their class predictions. The encoders remain frozen throughout. Because the classifiers operate in the shared label space, their outputs can be combined even when the encoders produce representations of different dimensions or were pretrained independently.

### 3.1 FACTORIZED CLASS-CONDITIONAL MODEL

Let $\pi_c = p(Y = c)$ be the class prior, and let $\mathcal{O} \subseteq \mathcal{M}$ index the modalities observed for an input. If their representations are conditionally independent given the class, Bayes' rule gives

$$p\Big(Y = c \mid \{z^{(m)}\}_{m \in \mathcal{O}}\Big) \propto \pi_c \prod_{m \in \mathcal{O}} p_m(z^{(m)} \mid Y = c). \tag{1}$$

The joint prediction thus depends on one class-conditional density per observed modality and a single class prior. These densities can be estimated separately in their respective representation spaces. Conditional independence is the modeling assumption behind the factorization.

### 3.2 Estimating Modality Likelihoods

We estimate the class-conditional likelihoods with a Gaussian discriminant model for each modality. The mean depends on the class, while the covariance is shared across classes:

$$Z^{(m)} \mid Y = c \sim \mathcal{N}(\mu_{m,c}, \Sigma_m). \tag{2}$$

For each modality, we estimate $\widehat{\mu}_{m,c}$ from the labeled representations of class $c$ for which that modality is available, and pool deviations from the class means to estimate $\widehat{\Sigma}_m$. Because the covariance may be difficult to invert with limited data, we use the regularized precision estimator

$$\widehat{\Omega}_m = d_m \left[ (N_m - 1)\widehat{\Sigma}_m + \operatorname{tr}(\widehat{\Sigma}_m) I_{d_m} \right]^{-1}, \tag{3}$$

where $N_m$ is the number of support examples with modality $m$ and $I_{d_m}$ is the identity matrix. This estimator follows Kubokawa & Srivastava (2008) and its use for frozen representations by Wang et al. (2024).

The shared covariance makes the class-dependent part of the Gaussian log-likelihood linear in $z$:

$$\widehat{\delta}_{m,c}(z) = z^\top \widehat{\Omega}_m \widehat{\mu}_{m,c} - \frac{1}{2} \widehat{\mu}_{m,c}^\top \widehat{\Omega}_m \widehat{\mu}_{m,c}. \tag{4}$$

Terms common to all classes have been omitted from this score.

### 3.3 Posterior Fusion and Prediction

We now obtain a class posterior from each fitted Gaussian model. Let $\widehat{\pi}_c = N_c/N$ be the empirical class prior, where $N_c$ is the number of support examples labeled $c$. Using this prior for every modality, Bayes' rule gives

$$\widehat{p}_m(c \mid z) = \frac{\widehat{\pi}_c \exp\Big(\widehat{\delta}_{m,c}(z)\Big)}{\sum_{k=1}^{C} \widehat{\pi}_k \exp\Big(\widehat{\delta}_{m,k}(z)\Big)}. \tag{5}$$

To combine these posteriors, we return to the factorization in Equation 1. Under its conditional-independence assumption, the fitted likelihood and posterior forms agree:

$$\begin{aligned} \widehat{p}\Big(Y = c \mid \{z^{(m)}\}_{m \in \mathcal{O}}\Big) &\propto \widehat{\pi}_c \prod_{m \in \mathcal{O}} \widehat{p}_m(z^{(m)} \mid Y = c) \\ &\propto \widehat{\pi}_c \prod_{m \in \mathcal{O}} \frac{\widehat{p}_m(c \mid z^{(m)})}{\widehat{\pi}_c}. \end{aligned} \tag{6}$$

The second form removes the prior already included in each modality posterior. Each modality's normalization is common to all classes and does not affect their comparison. Under conditional independence, this is the product rule for classifier combination (Kittler et al., 1998).

We implement the rule in log space. The evidence from modality $m$ for class $c$ is its change in log probability relative to the prior:

$$\widehat{e}_{m,c}(z) = \log \widehat{p}_m(c \mid z) - \log \widehat{\pi}_c. \tag{7}$$

Let $\mathcal{O}(x)$ denote the modalities available for input $x$. We combine their evidence and predict

$$\boxed{\widehat{y}(x) = \arg\max_c \left[ \log \widehat{\pi}_c + \alpha \sum_{m \in \mathcal{O}(x)} \widehat{e}_{m,c}(z^{(m)}) \right].} \tag{8}$$

At $\alpha = 1$, this is the most probable class under the fitted conditional-independence model. The shared coefficient $\alpha > 0$ sets the influence of the combined modality evidence relative to the prior;

we choose it on validation data without updating the encoders or Gaussian models. Missing modalities require no imputation or architectural modification. During adaptation, each Gaussian classifier uses all labeled examples for which its modality is available. At prediction time, we simply omit the evidence of any unavailable modality from the sum.

For multilabel tasks, we apply the same construction independently to each label as a binary problem and use its fused log-odds as the ranking score for average precision; the full formulation is given in Appendix A.3.

## 4 Experimental Setup

**Datasets.** We evaluate on six multimodal datasets with image and text information relevant for classification tasks. CUB-200-2011 (Wah et al., 2011) uses bird photographs as the image modality and crowdsourced natural-language descriptions (Reed et al., 2016) as the text modality; the task is 200-class bird species classification. MS-COCO (Lin et al., 2014) uses photographs and their human-written captions, with four classification tasks over animal, vehicle, indoor, and sports accessory categories. Visual Genome (Krishna et al., 2017) uses cropped image regions and their free-text region descriptions, with three tasks over objects, colors, and relations. CrisisMMD (Alam et al., 2018) uses images attached to tweets and the corresponding tweet text for disaster-type classification. MMIMDb (Arevalo et al., 2017) uses film posters and plot summaries for multi-label genre classification. MIMIC-CXR (Johnson et al., 2019) uses chest radiographs and their associated radiology reports for multi-label classification of clinical findings like edema and cardiomegaly. Complete dataset and task information is available in Appendix B.

**Encoders and Baselines.** We use three image encoders: CLIP ViT-B/16 (Radford et al., 2021), DINOv2-Base (Oquab et al., 2023), and a ViT-B/16 trained with class supervision (Dosovitskiy, 2020) and three text encoders: the CLIP text transformer, E5-Base (Wang et al., 2022), and all-MPNet-Base-v2 (Song et al., 2020). For both CLIP towers, we use the OpenCLIP LAION-2B checkpoint (Cherti et al., 2023). Our primary experiments use the eight combinations in which the image and text encoders were pretrained independently, and therefore do not assume a shared cross-modal representation space. The jointly pretrained CLIP image-text pair is evaluated separately in Appendix F.2.

We compare against three classes of baselines. First, gradient-free fusion methods combine predictions from the frozen modality encoders without fitting additional parameters. Second, learned fusion methods keep the encoders fixed but optimize a model that combines their representations. Third, encoder adaptation methods update the pretrained representations using either LoRA (Hu et al., 2021) or full backbone fine-tuning. Learned fusion and encoder adaptation baselines use gradient-based optimization, with model selection performed on the validation set. For our method, the fusion parameter $\alpha$ is selected on the validation set using a fixed one-dimensional search, requiring no gradient-based optimization.

**Evaluation Protocol.** Each method is fit on the training split, tuned on the validation split, and evaluated on the held-out test split. Single-label tasks are evaluated using accuracy while multi-label tasks are evaluated using macro one-vs-rest average precision. Our main evaluation considers three settings with limited data and/or missing modalities; results with complete supervision and both modalities available are reported in Appendix F.1.

[1] **Few-Shot Training.** We construct few-shot training sets with $k \in 1, 2, 4, 8, 16, 32, 64$ labeled examples per class and evaluate on the full test set. The validation set is subsampled to the same labeling budget so that model selection does not use additional labeled supervision.

[2] **Missingness in Test Data.** Training data retain both modalities, while validation and test examples are independently missing one modality with probability $r \in \{0.25, 0.5, 0.75\}$. When an example is masked, either its image or text modality is removed, but never both. We apply the same missingness rate to the validation and test sets so that model selection is performed under the same observation conditions encountered at inference time.

[3] **Few-Shot Training Using Missing Modalities.** For $k \in \{4, 16\}$, we independently remove one modality from training and validation examples with probability $r \in \{0.25, 0.5, 0.75\}$. Test examples retain both modalities.

# 5 RESULTS

## 5.1 FEW-SHOT TRAINING

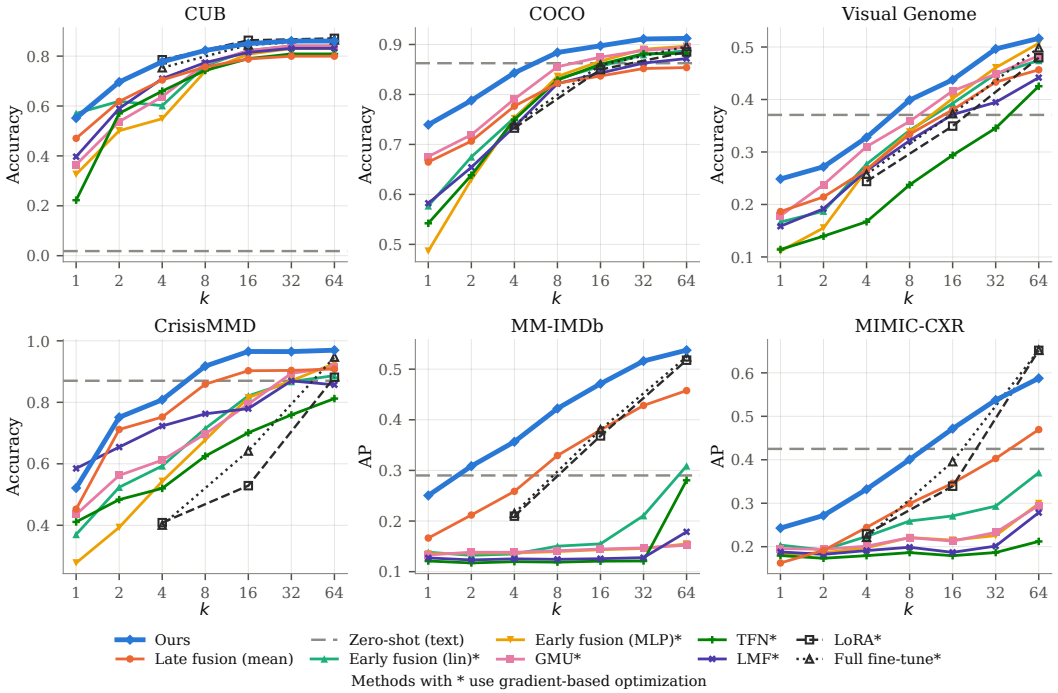



Figure 2: **Few-shot classification across six datasets, averaged over the eight encoder combinations.** Methods that require gradient-based optimization are marked with an asterisk (*).

**Baselines.** We compare our method with three kinds of methods: training-free fusion methods, learned fusion methods, and encoder adaptation. Learned fusion methods and encoder adaptation require gradient optimization. *For training-free fusion methods*, we consider two baselines: **(1) Zero-shot (text)** is a zero-shot classifier from the class names based on the pretrained text embedding space. **(2) Late fusion (mean)** (Kittler et al., 1998) uses the sum rule to combine the per-modality posteriors obtained from the same closed-form heads our method uses. *For learned fusion methods*, we consider five baselines: **(1) Early fusion (linear)** is a standard concatenation baseline where a linear classifier is trained on the concatenation of the image and text features. **(2) Early fusion (MLP)** passes the same concatenated representation to a two-layer perceptron, which can represent interactions between the two modalities that the linear head cannot. **(3) GMU** (Arevalo et al., 2017) learns a per-dimension gate that decides how much of each modality to admit, producing a single fused representation. **(4) TFN** (Zadeh et al., 2017) projects to a low-dimensional space and compute the outer product of the two representations, so that every pair of image and text dimensions is exposed to the classifier. **(5) LMF** (Liu et al., 2018) approximates that same outer product with low-rank factors, retaining the bimodal interactions at a small fraction of the parameters. Finally, we consider two *encoder adaptation* baselines: **(1) LoRA** (Hu et al., 2021) inserts low-rank adapters into a frozen backbone and trains only those; we apply rank-8 adapters to the attention projections of both encoders and train them jointly with the classification head. **(2) Full fine-tune** adapts all parameters of both encoders together with the classification head, making it the most expensive baseline we consider.

**Results.** Figure 2 compares our method with all baselines across six datasets, as the number of labeled examples increases. Across every dataset and shot count, our method consistently outperforms the gradient-free late fusion approach, with an average improvement of 7.47% over all settings. The gains are particularly pronounced on the multimodal multi-label tasks, including MM-IMDb and MIMIC-CXR, where the gap generally widens as more labeled examples become available. Notably, our gradient-free approach also outperforms methods that require gradient optimization, achieving

| Method | Gradient-Based Optimization | CUB Acc. | COCO Acc. | VG Acc. | CrisisMMD Acc. | MM-IMDb AP | MIMIC-CXR AP | Dataset Average |
|---|---|---|---|---|---|---|---|---|
| Zero-shot (text) | ✗ | 0.018 | 0.863 | 0.371 | 0.870 | 0.290 | 0.425 | 0.473 |
| Late fusion (mean) | ✗ | 0.789 ±0.001 | 0.797 ±0.007 | 0.380 ±0.010 | 0.903 ±0.001 | 0.380 ±0.003 | 0.346 ±0.004 | 0.606 |
| **Ours** | ✗ | **0.850** ±0.002 | **0.897** ±0.006 | **0.438** ±0.008 | **0.965** ±0.001 | **0.471** ±0.002 | **0.472** ±0.007 | **0.682** |
| Early fusion (linear) | ✓ | 0.822 ±0.004 | 0.856 ±0.018 | 0.393 ±0.011 | <u>0.821</u> ±0.025 | 0.155 ±0.005 | 0.271 ±0.005 | 0.553 |
| Early fusion (MLP) | ✓ | 0.808 ±0.001 | 0.867 ±0.013 | 0.402 ±0.014 | 0.815 ±0.047 | 0.143 ±0.001 | 0.215 ±0.002 | 0.542 |
| GMU | ✓ | 0.822 ±0.001 | <u>0.875</u> ±0.009 | <u>0.416</u> ±0.007 | 0.794 ±0.018 | 0.145 ±0.003 | 0.213 ±0.005 | 0.544 |
| TFN | ✓ | 0.790 ±0.003 | 0.860 ±0.003 | 0.294 ±0.045 | 0.701 ±0.094 | 0.121 ±0.002 | 0.179 ±0.004 | 0.491 |
| LMF | ✓ | 0.815 ±0.001 | 0.843 ±0.010 | 0.371 ±0.015 | 0.779 ±0.012 | 0.126 ±0.001 | 0.187 ±0.005 | 0.520 |
| LoRA | ✓ | <u>0.863</u> ±0.001 | 0.850 ±0.011 | 0.349 ±0.008 | 0.529 ±0.025 | 0.368 ±0.008 | 0.339 ±0.023 | 0.550 |
| Full fine-tune | ✓ | 0.845 ±0.001 | 0.859 ±0.004 | 0.373 ±0.005 | 0.642 ±0.013 | <u>0.380</u> ±0.010 | <u>0.396</u> ±0.014 | <u>0.582</u> |

Table 1: **16-shot classification on six datasets.** We evaluate on eight encoder combinations with three seeds per encoder. Values reported as means with standard deviations over 24 total seeds. Single-label datasets are scored by accuracy and multi-label datasets by macro one-vs-rest average precision. **Bold** denotes the best gradient-free result and <u>underline</u> the best gradient-optimized result. Zero-shot uses no labels and is deterministic.

the strongest overall performance on most datasets across much of the shot range ($36/42$ settings). These results indicate that the advantage of our method is not specific to a particular supervision level, but persists from extremely low-data settings through larger few-shot regimes.

Table 1 presents the 16-shot results for gradient-free and gradient-optimized baselines across six multimodal datasets. Our method consistently outperforms the gradient-free baselines on all six datasets, improving over the strongest gradient-free alternative by 3.4-9.1% across individual tasks and by $5.9\%$ on average. The gains are particularly pronounced on MM-IMDb, CrisisMMD, CUB, and Visual Genome, where our method improves by $9.1\%$, $6.2\%$, $6.1\%$, and $5.8\%$, respectively. Despite requiring no parameter optimization, our method also exceeds the best gradient-optimized baseline on five of the six datasets, with particularly large improvements on CrisisMMD, MM-IMDb, and MIMIC-CXR (average $10.4\%$). Overall, these results suggest that combining modalities through their class-conditional evidence provides an effective lightweight method for integrating multimodal information when limited training data is available.

### 5.2 Missingness in Test Data

**Additional Baselines.** The missingness setting admits two comparators that the few-shot setting does not. **(1) Modality dropout** is a standard remedy for missing inputs, in which each training example independently drops one modality with the same procedure applied at test time. The classifier then learns to predict from either modality alone rather than relying on both being present. We apply it to the strongest concatenation baseline, giving a variant of the early fusion (MLP) method that is explicitly trained for the condition it is evaluated under; **(2) ASIF** (Norelli et al., 2023): ASIF proposes to represent each item by its similarities to a set of complete image-text anchor pairs, rather than by its raw features, and to classify in that relative space. It is available here but not in the few-shot setting, because it requires at least fifty complete anchor pairs, which a $k$-shot support set does not supply. Under masking, we form its anchors from the training items that retain both modalities, and impute absent test features with the training mean of that modality.

**Results.** Figure 3 shows classification performance across varying rates of missingness in the evaluation set, representing the use case of using multimodal pipelines when both modalities are not available. Among gradient-free approaches, our method improves on late fusion by an average $10.96\%$ ($15/18$ settings) and on ASIF by an average $62.13\%$ ($18/18$). On CUB, our method exceeds every method that requires gradient-based optimization. On MM-IMDb, our method matches full encoder fine-tuning across the masked rates, attaining the performance of a fully adapted backbone without any gradient-based training. MIMIC-CXR is an exception, where fine-tuning leads all other less compute-intensive approaches by $0.185$-$0.237$.

We also compare with traditional approaches for handling missing modalities, including zero imputation, mean imputation, and modality dropout. As shown in Table 2, these strategies remain substantially less robust than our approach under increasing test-time missingness: at $r = 0.75$, our method exceeds zero imputation, mean imputation, and modality dropout by $6.3\%$, $4.8\%$, and $4.1\%$, respectively, while exhibiting a smaller overall drop from $r = 0$ to $r = 0.75$ ($7.6\%$ vs. $9.6$-$14.1\%$).

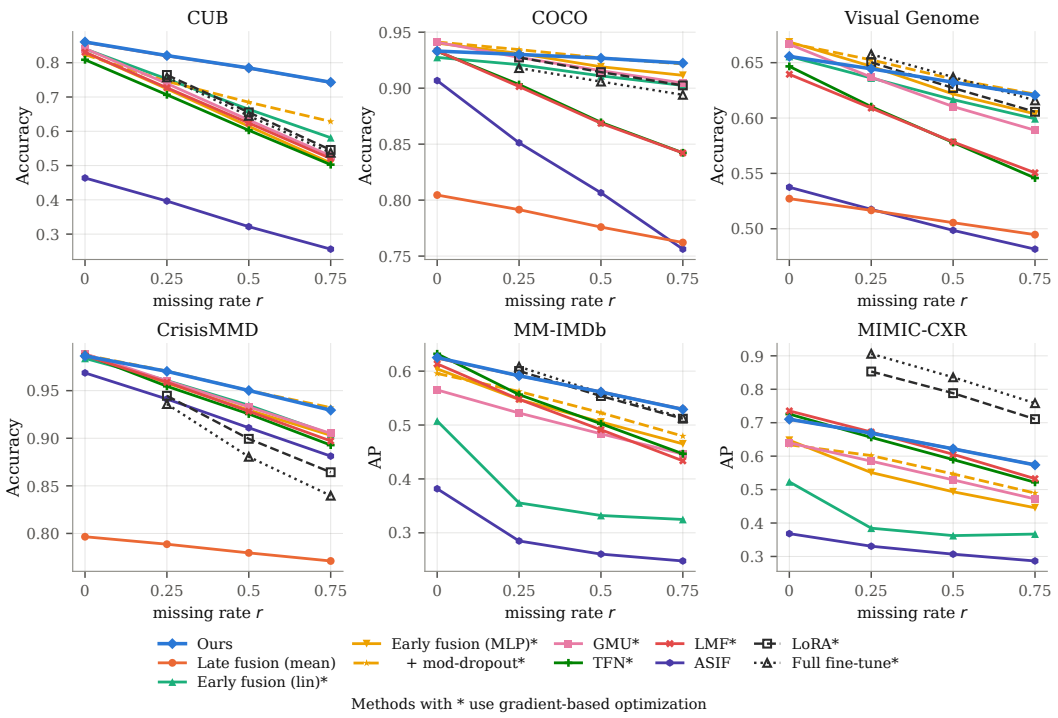



Figure 3: **Classification results with varying rates of missingness** $\{0, 0.25, 0.5, 0.75\}$ in the test data over six datasets and eight encoder combinations.

| Method | Imputes | $r=0$ | $r=0.25$ | $r=0.5$ | $r=0.75$ | $\Delta$ |
|---|---|---|---|---|---|---|
| Early fusion + zero imputation | ✓ | **0.798** | 0.748 | 0.701 | 0.657 | -0.141 |
| Early fusion + mean imputation | ✓ | **0.798** | 0.756 | 0.714 | 0.672 | -0.126 |
| Early fusion + modality dropout | ✓ | 0.775 | 0.745 | 0.711 | 0.679 | -0.096 |
| **Ours** | ✗ | 0.795 | **0.771** | **0.746** | **0.720** | **-0.076** |

Table 2: **Classification results with imputation strategies under test-time missingness.** Early fusion concatenates the two feature vectors and so must supply a value for a missing modality: zero imputation substitutes zeros, mean imputation substitutes the training mean of that modality, and modality dropout instead trains the model with modalities randomly removed. Our method does not use imputation since an absent modality simply contributes no evidence term. Results shown are the average across the six datasets and eight encoder combinations; $\Delta$ is the change from $r$=0 to $r$=0.75.

This indicates that explicitly modeling only the available modality evidence is more effective than imputing or training around absent inputs.

### 5.3 Few-Shot Training Using Missing Modalities

Figure 4 compares our method with baselines as the train/validation missing rate $r$ increases, holding the support set at only 16 labeled examples per class. Across this regime, our method is the strongest overall: it outperforms the best learned fusion baselines and encoder adaptation methods on all 18 dataset-$r$ settings with $r > 0$ $(18/18)$, despite the extra optimization those methods require. Averaged over all datasets, our method achieves the best performance in all six few-shot-missingness settings in Table 3, including 4-shot performance. FAME beats the strongest trained baseline in each setting by $8.0\%$ to $12.4\%$ points, and by $10.9\%$ at $k = 4$ and $8.1\%$ at $k = 16$ on average over the three missing rates. For example, averaged over $r \in \{0.25, 0.5, 0.75\}$, our method achieves $0.514$ versus $0.396$ for LoRA at $k = 4$, and $0.664$ versus $0.583$ for full fine-tuning at $k = 16$. Modality dropout does not close this gap under heavier missingness. These results highlight the sample efficiency of combining class-level evidence directly: under limited labeled supervision, our approach can exploit complementary multimodal information without fitting a fusion function from scarce data.

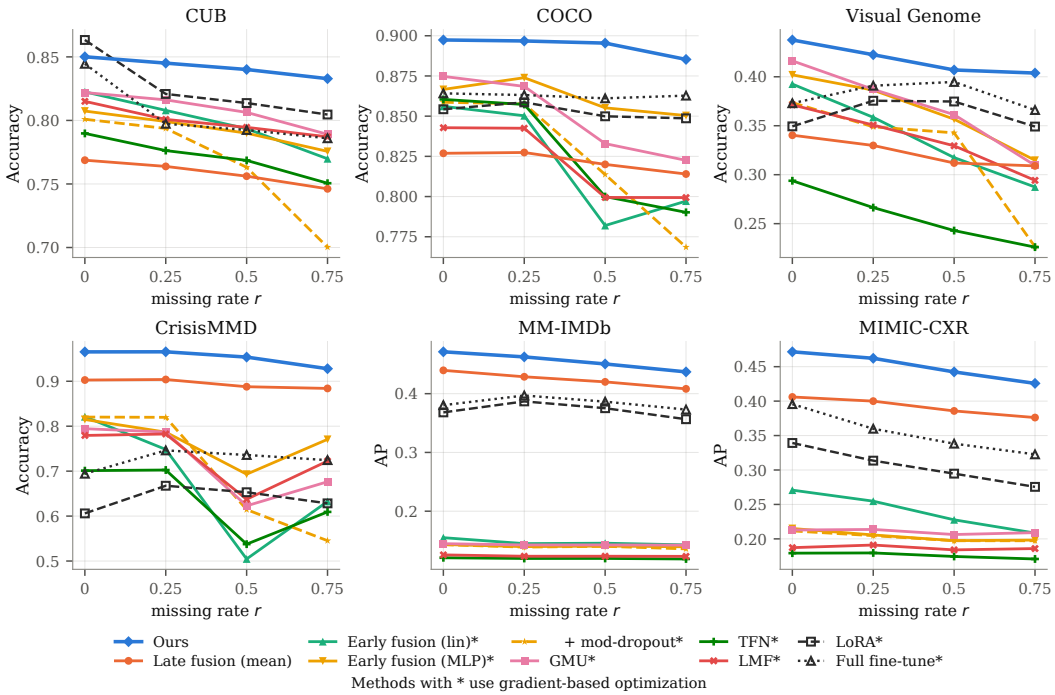



Figure 4: **16-shot classification with missing modality information across six datasets and eight encoder combinations.** We compare methods that require no training with methods that require training.

| | $k = 4$ | | | $k = 16$ | | |
|---|---|---|---|---|---|---|
| Method | $r$=0.25 | $r$=0.5 | $r$=0.75 | $r$=0.25 | $r$=0.5 | $r$=0.75 |
| Late fusion (mean) | 0.474 | 0.405 | 0.418 | 0.601 | 0.589 | 0.561 |
| **Ours** | **0.568** | **0.490** | **0.484** | **0.676** | **0.665** | **0.652** |
| Early fusion (lin) | 0.398 | 0.311 | 0.324 | 0.527 | 0.462 | 0.473 |
| Early fusion (MLP) | 0.406 | 0.320 | 0.296 | 0.532 | 0.506 | 0.508 |
| + modality dropout | 0.379 | 0.288 | 0.227 | 0.527 | 0.479 | 0.429 |
| GMU | 0.427 | 0.337 | 0.301 | 0.536 | 0.495 | 0.491 |
| TFN | 0.385 | 0.329 | 0.324 | 0.484 | 0.441 | 0.444 |
| LMF | 0.444 | 0.380 | 0.362 | 0.515 | 0.478 | 0.485 |
| LoRA | 0.417 | 0.393 | 0.377 | 0.570 | 0.560 | 0.544 |
| Full fine-tune | 0.405 | 0.385 | 0.376 | 0.592 | 0.585 | 0.572 |

Table 3: **Few-shot training with missing modalities.** Each training and validation item independently loses one modality with probability $r$, never both, on a $k$-shot pool. Values are means over the eight primary encoder combinations, six datasets, and three seeds per setting for the methods that train; single-label datasets are scored by accuracy and multi-label datasets by macro one-vs-rest average precision. **Bold** marks the best method in each column.

## 6 Conclusion

We study multimodal classification in regimes where supervision is limited and modality information may be incomplete. Our results show that strong multimodal performance does not necessarily require jointly adapting pretrained encoders or learning complex representation-level fusion. Instead, much of the useful structure can already be present within independently pretrained modality representations, with lightweight statistical adaptation sufficient to combine their predictive evidence effectively. This suggests a broader view of multimodal learning in data-constrained settings: rather than always learning richer cross-modal representations, it can be valuable to ask how far existing representations can be taken with simple, well-structured adaptation. We hope this perspective motivates further work on efficient multimodal methods that make better use of pretrained representations while reducing dependence on large paired datasets and costly optimization.

### AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on the manuscript. Otherwise, we have not used generative AI for any of the tasks that require disclosure. All figures and text were created and written by the authors. We used generative AI tools for fixing typos, phrasing, and formatting of results tables. We also used generative AI to edit our code repository, particularly for removing MIMIC dataset leakage in the code repository, as per the PhysioNet credentialized user agreement. We have reviewed all AI-generated work and carefully incorporated the limited suggestions provided by the tools. We take complete responsibility for the final content of this work.

### ETHICS STATEMENT

Our experiments use existing image-text datasets, including MIMIC-CXR, which is available under a credentialed data use agreement. We do not reproduce radiographs or reports from MIMIC-CXR in this paper. Its clinical finding labels are derived from the same reports used as text inputs. Performance on this task therefore measures prediction of report-derived labels, not agreement with an independent clinical assessment or the ability to diagnose from an image alone.

The method inherits limitations of its datasets and pretrained encoders. In particular, performance under the simulated missingness patterns studied here does not establish performance when modality availability depends on patient characteristics or clinical circumstances. We do not assess subgroup error rates or prospective clinical outcomes. The MIMIC-CXR results should not be taken as evidence that this method is suitable for clinical decision-making without independent labels, evaluation across patient groups and missingness patterns, and clinical oversight.

### REPRODUCIBILITY STATEMENT

We provide a link to the code repository and are committed to ensuring reproducibility. Appendices A- B specify the method implementation, dataset splits, encoder checkpoints, baseline settings, and evaluation protocols needed to reproduce our results.

## REFERENCES

Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

Yichao Cai, Yuhang Liu, Erdun Gao, Tianjiao Jiang, Zhen Zhang, Anton van den Hengel, and Prof Javen Qinfeng Shi. On the value of cross-modal misalignment in multimodal representation learning. *Advances in Neural Information Processing Systems*, 38:145859–145911, 2026.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE transactions on signal processing*, 58(10):5016–5029, 2010.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20029–20038. IEEE, 2023.

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7 (2):179–188, 1936.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 746–754, 2023.

Trevor Hastie and Robert Tibshirani. Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):155–176, 1996.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C 2 kd: Bridging the modality gap for cross-modal knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015. IEEE, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998. doi: 10.1109/34.667881.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Tatsuya Kubokawa and Muni S Srivastava. Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data. *Journal of multivariate Analysis*, 99(9): 1906–1928, 2008.

Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2247–2256, 2018.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 2302–2310, 2021.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Icml*, volume 11, pp. 689–696, 2011.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodola, and Francesco Locatello. Asif: Coupled data turns unimodal models to multimodal without training. *Advances in Neural Information Processing Systems*, 36:15303–15319, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2017.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736. IEEE, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15878–15887. IEEE, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1828–1838, 2020a.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12692–12702. IEEE, 2020b.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation, 2024. URL https://arxiv.org/abs/2402.04087.

Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20039–20049. IEEE, 2023.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017. URL https://arxiv.org/abs/1707.07250.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 2418–2428, 2022.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

## APPENDIX OUTLINE

This Appendix is organized as follows:

- In Appendix A, we provide a pseudocode block to illustrate the implementation of FAME and explain the nuances of degenerate cases.
- In Appendix B, we include details on datasets, tasks, and implementation with regards to training protocols and baselines.
- In Appendix C, we include a computational breakdown of our experiments, including timing and efficiency analyses of our method.
- In Appendix D, we provide method-oriented ablations to understand how components of FAME combine to make an effective method. We analyze across the different data-constrained settings.
- In Appendix E, we provide derivations and theoretical properties of FAME not discussed in the main text.
- In Appendix F, we provide additional results such as performance on the complete training data, with jointly pretrained CLIP encoders, and breakdown by encoder pair.
- In Appendix G, we include the proper attribution to images used in Figure 1.

# A Details of the Method

Algorithm 1 summarizes the multiclass inference procedure for FAME; the multilabel extension is given in Appendix A.3. All modality-specific statistics are estimated directly from the labeled support set (training data); the only selected hyperparameter is the scalar fusion coefficient $\alpha$.

## A.1 Pseudocode

**Algorithm 1** PyTorch-like pseudocode for multiclass FAME.

```
# Input:
#   - Z: support representations for M modalities
#   - O: support observation mask
#   - Y: support labels
#   - Z_val, O_val, Y_val: validation data
#   - Z_test, O_test: test data
#   - alphas: candidate fusion coefficients
# Output:
#   - y_hat: predicted test labels

def method(Z, O, Y, Z_val, O_val, Y_val, Z_test, O_test, alphas):

    # 1. Estimate the shared class prior.
    log_pi = estimate_log_prior(Y)

    # 2. Fit one Gaussian discriminant classifier per modality.
    classifiers = []
    for m in range(M):
        Xm = Z[m][O[:, m] == 1]
        Ym = Y[O[:, m] == 1]
        W_m, b_m = fit_gda(Xm, Ym)
        classifiers.append((W_m, b_m))

    def evidence(Zs, Os):
        E = 0
        for m in range(M):
            W_m, b_m = classifiers[m]

            # Modality-specific posterior.
            delta_m = Zs[m] @ W_m + b_m
            log_p_m = log_softmax(delta_m + log_pi, dim=-1)

            # Remove the shared class prior and ignore missing modalities.
            e_m = log_p_m - log_pi
            E += Os[:, m].unsqueeze(-1) * e_m

        return E

    def score(Zs, Os, alpha):
        return log_pi + alpha * evidence(Zs, Os)

    # 3. Select the fusion coefficient on validation data.
    alpha = max(
        alphas,
        key=lambda a: metric(score(Z_val, O_val, a), Y_val)
    )

    # 4. Predict from the combined class evidence.
    return score(Z_test, O_test, alpha).argmax(-1)
```

## A.2 Implementation Details

**Class prior.** For multiclass tasks, we estimate a single class prior from the full labeled support set (in practice, this is the training data) $\hat{\pi}_c = \frac{N_c}{N}$, where $N_c$ denotes the number of support examples in class $c$. This prior is shared across modalities.

**Modality-specific GDA estimation.** For each modality $m$, we estimate Gaussian statistics using only support examples for which that modality is observed. Define

$$\mathcal{I}_m = \{i : O_{i,m} = 1\}, \qquad N_m = |\mathcal{I}_m|, \qquad N_{m,c} = \sum_{i \in \mathcal{I}_m} \mathbb{1}[y_i = c].$$

For each class represented in modality $m$, the empirical class mean is

$$\hat{\mu}_{m,c} = \frac{1}{N_{m,c}} \sum_{\substack{i \in \mathcal{I}_m \\ y_i = c}} z_i^{(m)}.$$

If $N_{m,c} = 0$, we instead use the pooled modality mean,

$$\hat{\mu}_{m,c} = \frac{1}{N_m} \sum_{i \in \mathcal{I}_m} z_i^{(m)},$$

which yields a finite, approximately neutral contribution from that modality rather than an undefined class score.

**Covariance and precision estimation.** For each observed support example, let $r_i^{(m)} = z_i^{(m)} - \hat{\mu}_{m,y_i}$ denote its within-class residual. We estimate the pooled within-class covariance as

$$\hat{\Sigma}_m = \frac{1}{\max\{N_m - C,\, 1\}} \sum_{i \in \mathcal{I}_m} r_i^{(m)} r_i^{(m)\top}.$$

The corresponding regularized precision matrix is

$$\hat{\Omega}_m = d_m \left( (N_m - 1)\hat{\Sigma}_m + \mathrm{tr}(\hat{\Sigma}_m) I_{d_m} \right)^{-1},$$

as defined in Eq. 3. The discriminant parameters are then

$$W_{m,\cdot c} = \hat{\Omega}_m \hat{\mu}_{m,c}, \qquad b_{m,c} = -\frac{1}{2} \hat{\mu}_{m,c}^\top \hat{\Omega}_m \hat{\mu}_{m,c}.$$

In practice, we do not explicitly form $\hat{\Omega}_m$. We symmetrize the regularized covariance matrix, compute a single Cholesky factorization, and solve simultaneously for all class means.

**Degenerate covariance estimates.** At one shot per class, every within-class residual is zero and $\hat{\Sigma}_m$ vanishes. In this case, we construct the algorithm such that we can use the total covariance of the observed support representations,

$$\hat{\Sigma}_m^{\mathrm{tot}} = \frac{1}{\max\{N_m - 1,\, 1\}} \sum_{i \in \mathcal{I}_m} \left( z_i^{(m)} - \bar{z}^{(m)} \right) \left( z_i^{(m)} - \bar{z}^{(m)} \right)^\top,$$

where

$$\bar{z}^{(m)} = \frac{1}{N_m} \sum_{i \in \mathcal{I}_m} z_i^{(m)}.$$

If the total covariance is also degenerate, we use $I_{d_m}$, reducing the discriminant to an isotropic nearest-mean rule.

**Class evidence and missing modalities.** For an observed modality $m$, the modality-specific posterior is

$$\log \hat{p}_m(c \mid z^{(m)}) = \mathrm{logsoftmax}_c \left( \hat{\delta}_{m,c}(z^{(m)}) + \log \hat{\pi}_c \right).$$

We define class evidence by removing the common prior,

$$e_{m,c} = \log \hat{p}_m(c \mid z^{(m)}) - \log \hat{\pi}_c.$$

Missing modalities contribute zero evidence. Thus, for the set of observed modalities $\mathcal{O}(x)$, prediction is

$$\hat{y} = \arg\max_c \left\{ \log \hat{\pi}_c + \alpha \sum_{m \in \mathcal{O}(x)} e_{m,c} \right\}.$$

**Fusion coefficient.** The modality-specific Gaussian classifiers are determined entirely by the support set. The only selected hyperparameter is the scalar fusion coefficient $\alpha$, chosen on the validation split from

$$\alpha \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\right\}.$$

We select the value maximizing the validation metric for the corresponding task. No gradient-based optimization is used.

### A.3 Multilabel Extension

For a task with $L$ labels, we treat each label $\ell$ as a separate binary problem. For each modality $m$ and label $\ell$, we estimate a mean for the positive examples and a mean for the negative examples among the $N_m$ support examples in which $m$ is observed. We pool the within-group scatter and divide by $\max\{N_m - 2, 1\}$; the label-specific regularized precision uses the same $N_m - 1$ factor as in the multiclass estimator above. We apply the same fallback for a degenerate covariance estimate. If one binary group is absent among the observed support examples for a modality, that modality contributes zero evidence for the corresponding label.

Let $\hat{\delta}_{m,\ell,+}$ and $\hat{\delta}_{m,\ell,-}$ be the positive and negative Gaussian discriminants. We estimate the label prior from the full support set with Laplace smoothing, $\hat{\pi}_\ell = (N_{\ell,+} + 1)/(N + 2)$, where $N_{\ell,+}$ is the number of support examples positive for label $\ell$. The fused score is the prior log-odds plus the evidence from observed modalities:

$$s_\ell(x) = \log \frac{\hat{\pi}_\ell}{1 - \hat{\pi}_\ell} + \alpha \sum_{m \in \mathcal{O}(x)} \left[\hat{\delta}_{m,\ell,+}(z^{(m)}) - \hat{\delta}_{m,\ell,-}(z^{(m)})\right].$$

We rank examples independently for each label using $s_\ell(x)$ to compute one-vs-rest average precision. There is no softmax or argmax across labels.

## B Experimental Details

### B.1 Datasets and Tasks

We evaluate on six publicly available multimodal datasets spanning eleven classification tasks. Each example consists of an image and naturally occurring text associated with the same underlying instance, such as a caption, tweet, plot summary, or radiology report. The image-text pairings are provided by the original datasets; we do not construct text inputs from class names or prompts. Table 4 summarizes the dataset splits and label spaces, Table 5 specifies the prediction targets, and Table 6 shows representative image-text pairs.

Some datasets associate multiple texts with the same image. CUB-200-2011 provides ten descriptions per image and MS-COCO provides five captions per image. For these datasets, we encode each text independently and use the normalized mean of the resulting representations as the text representation. Datasets with a single associated text are encoded directly. MIMIC-CXR-JPG contains 377,110 radiographs from 227,835 studies and 65,379 patients. We retain frontal AP and PA views, yielding 243,334 images from 218,139 studies. Each image is paired with its study-level free-text radiology report and associated task labels.

The source of the prediction labels also differs across datasets. For MS-COCO and Visual Genome, labels are obtained from the datasets' object, attribute, and relation annotations and therefore need not be explicitly mentioned in the associated text.

### B.2 Implementation Details

**Encoders.** Table 7 lists the exact checkpoints. Both CLIP towers come from the OpenCLIP `ViT-B-16` LAION-2B release (`laion2b_s34b_b88k`) (Cherti et al., 2023). Crossing three image encoders with three text encoders gives nine configurations; the eight in which the two towers were pretrained independently form our primary setting, and the jointly pretrained CLIP-CLIP pair is reported separately in Appendix F.2. All representations are extracted once from the frozen encoders, cached, and $L_2$-normalized before any classifier is fit. Encoders are never updated except in the encoder-adaptation baselines.

| Dataset | Task | Train | Val. | Test | Total | Labels |
|---|---|---|---|---|---|---|
| CUB-200-2011 | Species | 5,000 | 994 | 5,794 | 11,788 | 200 species |
| MS-COCO | Animal | 13,419 | 267 | 570 | 14,256 | 4 classes |
| | Indoor object | 21,631 | 431 | 935 | 22,997 | 4 classes |
| | Sports accessory | 14,438 | 314 | 605 | 15,357 | 4 classes |
| | Vehicle | 11,103 | 238 | 480 | 11,821 | 3 classes |
| Visual Genome | Object | 6,578 | 884 | 812 | 8,274 | 8 classes |
| | Color | 13,939 | 1,830 | 1,806 | 17,575 | 7 classes |
| | Relation | 5,461 | 662 | 656 | 6,779 | 7 classes |
| CrisisMMD | Disaster type | 13,608 | 2,237 | 2,237 | 18,082 | 4 classes |
| MM-IMDb | Genre | 15,540 | 2,606 | 7,794 | 25,940 | 23 genres |
| MIMIC-CXR | Clinical findings | 237,972 | 1,959 | 3,403 | 243,334 | 12 findings |

Table 4: Dataset statistics for the eleven classification tasks. Splits are fixed and shared across all methods. MM-IMDb and MIMIC-CXR are multilabel tasks; all remaining tasks are multiclass. MIMIC-CXR contains 12 independently evaluated clinical findings. MM-IMDb examples contain 2.48 genre labels on average, with at most 9 labels per example.

| Dataset | Task | Label set |
|---|---|---|
| CUB-200-2011 | Species | 200 bird species |
| MS-COCO | Animal | bird, cat, dog, horse |
| | Indoor object | cell phone, clock, dining table, sink |
| | Sports accessory | sports ball, surfboard, tie, umbrella |
| | Vehicle | bus, train, truck |
| Visual Genome | Object | man, woman, shirt, head, leg, trouser, hair, tree |
| | Color | white, black, blue, brown, red, green, grey |
| | Relation | wear, in, behind, under, next to, about, sit |
| CrisisMMD | Disaster type | hurricane, earthquake, wildfire, flood |
| MM-IMDb | Genre | 23 film genres, including Drama, Comedy, Romance, Thriller, and Crime |
| MIMIC-CXR | Clinical findings | Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Support Devices |

Table 5: **Prediction targets for the evaluation tasks.** The four MS-COCO tasks restrict prediction to selected object categories within the corresponding COCO supercategory. Visual Genome tasks are defined over annotated regions, so the image input is the corresponding region crop rather than the full image. Each task is trained and evaluated independently.

**Our method.** All modality-specific statistics are computed in closed form, so the method has no learning rate, no epochs, no initialization, and no random seed: two runs on the same support set produce identical output. We use the full pooled within-class covariance with the shrinkage of Equation 3 applied as published, i.e. with a unit coefficient on the $\mathrm{tr}(\widehat{\Sigma}_m)I$ term. The single fusion coefficient $\alpha$ is selected on the validation split over the seven-point logarithmic grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$, shared across modalities, with the prior coefficient fixed at 1. Selection uses the same metric and the same data conditions as the final evaluation, and the test split is scored exactly once. The precision matrix is never formed explicitly: the bracketed matrix in Equation 3 is symmetrized, Cholesky-factored once per modality, and solved against the stacked class means.

Because the method is deterministic, we do not average it over seeds; reported variability for our method comes from resampling the data (the few-shot support sets and the missingness masks), not from re-initialization.

Table 6: Representative image-text pairs from the evaluation tasks. Examples are drawn from the corresponding test splits. For CUB-200-2011 and MS-COCO, which associate multiple descriptions with each image, one description is shown for readability. Visual Genome examples use the annotated region crop provided to the model. Text is truncated for display where necessary; models receive the complete input. The MIMIC-CXR example is omitted in compliance with the *PhysioNet Credentialed Health Data Use Agreement.*

| Dataset | Task | Image | Text | Label |
|---|---|---|---|---|
| CUB-200-2011 | Species |  | "a greyish-blue colored bird with a white-grey face, and big blue feet." | Black-footed Albatross |
| MS-COCO | Animal |  | "Three birds walking around a dry grass field." | bird |
| | Indoor object |  | "A tan toilet and sink combination in a small room." | sink |
| | Sports accessory |  | "A lady in a crowd holding an umbrella with a British flag design." | umbrella |
| | Vehicle |  | "A bus traveling down a curvy road behind a black car." | bus |
| Visual Genome | Object |  | "man wearing a pair of sun glasses" | man |
| | Color |  | "Brown wooden box" | brown |
| | Relation |  | "man wearing gray striped shirt" | wear |
| CrisisMMD | Disaster type |  | "Hurricane Maria batters Puerto Rico as a Cat 4; monitoring the track as it nears the continental US next week" | hurricane |
| MM-IMDb | Genre |  | "Allan Quatermain has been recruited to lead an expedition in search of a fabled treasure, deep within Africa. He must avoid hidden dangers." | Action, Adventure |

**Training-free fusion baselines.** The sum and product rules have no fitted parameters. The weighted posterior average searches its mixing weight over 21 values evenly spaced on $[0, 1]$. ASIF (Norelli et al., 2023) searches its number of anchors over $\{200, 800\}$ and its exponent over $\{4, 8\}$ on validation; it requires at least 50 complete image-text anchor pairs and is therefore omitted from the few-shot and few-shot-missingness settings, where that condition cannot be met. The closed-form early-fusion control fits the same Gaussian head on the concatenated representation and searches its ridge coefficient over $\{0.1, 0.3, 1, 3, 10, 30, 100\}$, since the concatenated covariance is the worst-conditioned estimate in the suite.

| Modality | Encoder | Checkpoint | Dim. | Context |
|---|---|---|---|---|
| Image | CLIP ViT-B/16 | `open_clip ViT-B-16` (LAION-2B) | 512 | – |
| | DINOv2-Base | `facebook/dinov2-base` | 768 | – |
| | ViT-B/16 | `google/vit-base-patch16-224` | 768 | – |
| Text | CLIP text | `open_clip ViT-B-16` (LAION-2B) | 512 | 77 tokens |
| | E5-Base | `intfloat/e5-base-v2` | 768 | 512 tokens |
| | all-MPNet-Base-v2 | `sentence-transformers/all-mpnet-base-v2` | 768 | 384 tokens |

Table 7: **Encoder checkpoints.** E5-Base is queried with the required ``query: '' prefix. Text encoders differ substantially in context length, which is relevant for MM-IMDb plot summaries and MIMIC-CXR reports.

**Learned fusion baselines.** Linear and MLP early fusion, GMU (Arevalo et al., 2017), TFN (Zadeh et al., 2017), and LMF (Liu et al., 2018) all operate on frozen representations. Hidden width is $128$ for the MLP, GMU, and LMF, and $32$ for TFN, whose outer-product interaction makes a wider layer prohibitive; LMF uses rank $4$. All are optimized with Adam, weight decay $10^{-4}$, batch size $256$, for at most $60$ epochs, with the learning rate selected on validation from $\{10^{-3}, 10^{-4}\}$. Early stopping monitors the validation *task metric* rather than the loss, with patience $8$. Each configuration is run with three seeds $(0, 1, 2)$; we report the mean and standard deviation across seeds, and each seed is scored separately rather than having its logits averaged.

**Encoder adaptation baselines.** LoRA (Hu et al., 2021) inserts rank-$8$ adapters with scaling $\alpha_{\text{LoRA}} = 16$ into the attention projections of both towers, using the target modules appropriate to each architecture (`query`/`value` for the Hugging Face towers, `c_fc`/`c_proj` for CLIP), with the base weights frozen and the $B$ matrix zero-initialized. Full fine-tuning updates all encoder parameters. Both are optimized with AdamW, weight decay $0.01$, batch size $32$, and a cosine-annealed schedule, for at most $30$ epochs with patience $5$ on the validation metric. Learning rates are set per parameter group: the classification head uses $10^{-3}$, while the encoder uses $10^{-5}$ for full fine-tuning and $10^{-4}$ for LoRA. Each is run with three seeds $(0, 1, 2)$.

**Linear probes.** The discriminative reference heads are $L_2$-regularized logistic regressions fit on the same frozen representations, with the inverse regularization strength selected on validation from $\{0.01, 1\}$ in the full-data setting and from $\{0.01, 0.1, 1, 10\}$ in the few-shot setting, where an unregularized probe separates the support set perfectly and generalizes at chance.

**Evaluation and model selection.** Every method is fit on the training split, has its hyperparameters selected on the validation split, and is scored once on the test split. Model selection uses accuracy for multiclass tasks and macro one-vs-rest average precision for multilabel tasks; we additionally report balanced accuracy and macro AUROC. Selection and evaluation always use the same metric definition, and no method sees the test split during selection.

**Few-shot protocol.** We sample $k \in \{1, 2, 4, 8, 16, 32, 64\}$ labeled examples per class, with three resampling seeds $(0, 1, 2)$ per budget, and evaluate on the full test split. The validation split is subsampled to the *same* $k$-per-class budget as the support set. Selecting hyperparameters on a full validation split while fitting on $k$ examples is not a few-shot protocol and would specifically favor methods whose hyperparameters are chosen on validation, including ours. For multilabel tasks the support set is grown until every label has at least $k$ positives, so that no label is left unestimable.

**Missingness protocol.** Each affected example loses one modality with probability $r \in \{0, 0.25, 0.5, 0.75\}$, never both, with three mask seeds $(0, 1, 2)$. We consider three regimes: missing supervision (training and validation incomplete, test complete), missing supervision under a few-shot budget ($k \in \{4, 16\}$), and missing observations (training complete, validation and test incomplete). In the last regime the validation and test masks use the same rate, so model selection occurs under the observation pattern seen at test time. Methods that require a fixed-width input

| Method | Median fit time per task (s) | | | |
|---|---|---|---|---|
| | Full data | $k$=4 | $k$=16 | Missing test |
| *No gradient-based training* | | | | |
| **Ours** | **0.49** | 0.27 | 0.28 | **0.50** |
| Late fusion (mean) | 0.49 | 0.27 | 0.28 | 0.50 |
| ASIF | 5.62 | — | — | 5.94 |
| *Trained fusion head* | | | | |
| Early fusion (lin) | 1.75 | **0.09** | 0.15 | 1.97 |
| Early fusion (MLP) | 2.21 | **0.05** | **0.08** | 2.02 |
| GMU | 3.11 | 0.06 | 0.11 | 2.96 |
| TFN | 3.13 | 0.06 | 0.08 | 2.91 |
| LMF | 3.12 | 0.08 | 0.13 | 3.10 |
| *Encoder adaptation* | | | | |
| LoRA | 592.1 | 11.2 | 18.9 | 570.8 |
| Full fine-tune | 728.5 | 11.0 | 19.9 | 757.2 |

Table 8: Median fitting time for one (dataset, task, encoder pair) configuration on a single GPU. For frozen-feature methods, times exclude the feature-extraction step used to construct the pretrained representations. Missing entries indicate settings in which a baseline is not evaluated. ASIF requires at least 50 complete image–text anchor pairs and is therefore omitted from the few-shot regimes. Bold indicates the fastest method in each column.

are given the standard remedies (zero imputation, mean imputation, and modality dropout during training) so the comparison is against baselines equipped for missingness rather than naive ones.

**Multilabel tasks.** For MM-IMDb and MIMIC-CXR, we treat each label as a binary prediction target. The late-fusion rules combine the corresponding positive-class probabilities from the modality-specific Gaussian heads; the product rule also combines the negative-class probabilities before normalizing the two outcomes. The closed-form early-fusion control fits a binary Gaussian head per label on the concatenated representation, while linear probes fit separate $L_2$-regularized binary logistic regressions with a shared regularization strength. ASIF forms a prototype for each label from training texts carrying that label and scores examples by their image- and text-side similarity to it.

For learned fusion and encoder adaptation, the classifier produces a logit $h_{i,c}$ for each example $i$ and label $c$. These methods minimize binary cross-entropy averaged over examples and labels:

$$\mathcal{L} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[y_{i,c}\log\sigma(h_{i,c}) + (1-y_{i,c})\log\big(1-\sigma(h_{i,c})\big)\right],$$

where $\sigma$ is the sigmoid function. Each label's positive and negative examples therefore contribute to its loss independently of the other labels assigned to the same example.

## C Computational Breakdown

We evaluate efficiency along three dimensions: the wall-clock cost of adapting to a new task, the amount of task-specific state introduced by each method, and the cost of inference. We additionally report the computational resources used for the full experimental study.

**Fitting time.** Table 8 reports the median wall-clock fitting time for one (dataset, task, encoder pair) configuration. Each job was allocated a single GPU from the heterogeneous compute pool described below. We therefore focus on the large differences between computational regimes rather than small differences in absolute runtime.

With full training data, our method fits in $0.49$ seconds, compared with $592$ seconds for LoRA and $729$ seconds for full fine-tuning, corresponding to speedups of $1{,}208\times$ and $1{,}487\times$, respectively. Relative to learned fusion heads operating on the same frozen representations, the difference is smaller, ranging from $3.6\times$ to $6.4\times$.

| Method | Task-specific values | Adaptation |
|---|---|---|
| **Ours** | **6,153** | closed form |
| Late fusion (mean) | 6,148 | closed form |
| Early fusion (lin) | 8,967 | gradient-based |
| TFN | 53,576 | gradient-based |
| Early fusion (MLP) | 197,252 | gradient-based |
| LMF | 200,996 | gradient-based |
| GMU | 394,116 | gradient-based |
| LoRA | 787,076 | gradient-based |
| Full fine-tune | 196,259,972 | gradient-based |

Table 9: Task-specific quantities for a representative four-class task with two 768-dimensional encoders. Our method estimates its 6,153 stored values in closed form and does not update either pretrained encoder. Gradient-based methods report the number of trainable task-specific parameters. ASIF is omitted because its task-specific storage is determined by the number and dimension of stored anchors rather than a directly comparable parameter count.

In the few-shot regime, trained fusion heads are slightly faster than our method: at $k{=}4$ and $k{=}16$, they fit in $0.05$–$0.15$ seconds compared with approximately $0.27$ seconds for ours. Optimization over a very small support set is inexpensive, whereas our method forms a covariance matrix for each modality and performs a $d_m \times d_m$ factorization. Covariance estimation scales with the number of support examples, while the factorization contributes an $O(d_m^3)$ term that does not decrease with $k$. Consequently, all frozen-feature approaches remain sub-second in this regime; the larger computational separation is relative to encoder adaptation, which remains approximately $40\times$–$70\times$ slower.

**Task-specific parameters.** Table 9 compares the amount of task-specific state introduced by each method. Our method does not update the pretrained encoders or learn a fusion network. Instead, it estimates one linear discriminant per class and modality and selects the scalar $\alpha$ from a seven-point validation grid.

For $C$ classes and modality dimensions $d_m$, the stored task-specific quantities total

$$\sum_m (Cd_m + C) + 1.$$

For a four-class task with two 768-dimensional encoders, this corresponds to 6,153 values. In contrast, LoRA introduces approximately 0.8M task-specific parameters, while full fine-tuning updates the full 196M parameters of the encoder pair.

**Feature reuse.** Freezing the encoders allows representations to be extracted once and reused across downstream tasks. The same cached representations support our method and the other frozen-feature baselines without re-running or updating the encoder for each task. LoRA and full fine-tuning instead produce task-specific encoder weights, requiring separate adaptation and feature computation for each downstream task.

**Inference cost.** After fitting, our classifier requires

$$\sum_m Cd_m$$

multiply-accumulates per example, excluding encoder evaluation. This is 6,144 operations for the representative four-class, two-modality setting, matching the leading-order cost of a linear fusion head and requiring approximately $32\times$ fewer operations than the width-128 MLP fusion head. Encoder evaluation remains the dominant shared component of inference. When a modality is unavailable at test time, our method simply omits its evidence term, reducing classifier-side computation without requiring reconstruction or imputation.

**Computational resources.** Each experimental job was allocated one GPU, 8 CPU cores, and 64 GB of system memory. Experiments were run on a heterogeneous GPU pool consisting of

| Variant | Score | $\Delta$ |
|---|---|---|
| *Fusion and weighting* | | |
| **Ours** (shared $\alpha$) | **0.807** | — |
| $\alpha = 1$ | 0.756 | $-0.051$ |
| posterior mean | 0.708 | $-0.099$ |
| without prior anchoring | 0.706 | $-0.101$ |
| without prior anchoring, $\alpha = 1$ | 0.683 | $-0.124$ |
| *Gaussian estimator* | | |
| diagonal covariance | 0.742 | $-0.065$ |
| isotropic covariance | 0.739 | $-0.068$ |
| without the $(N - 1)$ factor | 0.744 | $-0.063$ |
| *Logistic head* | | |
| logistic + shared $\alpha$ | 0.802 | $-0.005$ |
| logistic + $\alpha = 1$ | 0.804 | $-0.003$ |
| logistic + posterior mean | 0.803 | $-0.004$ |
| logistic without prior anchoring | 0.801 | $-0.006$ |

Table 10: Full training data ablations, averaged over eight independently pretrained encoder pairs and eleven tasks. $\Delta$ is measured relative to the complete method.

NVIDIA A100 (80 GB), RTX A6000 (48 GB), L40S (48 GB), and H200 (141 GB) GPUs. By allocated GPU-time, $49.7\%$ of the study ran on RTX A6000 GPUs, $44.7\%$ on A100 GPUs, $5.5\%$ on L40S GPUs, and $0.2\%$ on H200 GPUs. Including all methods, encoder pairs, tasks, regimes, ablations, and superseded runs, the complete experimental campaign comprised $10{,}708$ GPU jobs and $1{,}232$ GPU-hours of allocated wall-clock time. Of these, $876$ GPU-hours correspond to model fitting; the remainder primarily comprises data loading and encoder forward passes used to construct representations. Encoder-adaptation baselines account for $823$ hours ($94\%$), compared with $21$ hours for trained fusion heads and $5.5$ hours for our method across all evaluated configurations, including ablations and validation over $\alpha$. The computational difference therefore compounds over experimental grids spanning multiple encoder pairs, tasks, shot counts, and test-time missing-modality settings.

# D Ablations

We examine three components of FAME: the covariance structure used by the Gaussian discriminant, the rule used to combine modality evidence, and the coefficient that balances this evidence against the class prior. We also replace the Gaussian classifier with a logistic head to separate the effect of the per-modality evidence model from the fusion rule. Unless otherwise noted, results are averaged over eight independently pretrained encoder pairs and eleven tasks.

## D.1 Full Training Data

With the full training set, the covariance model and the evidence combination both contribute to performance. Replacing the full covariance with a diagonal or isotropic estimate reduces the aggregate score by $0.065$ and $0.068$, respectively. The diagonal and isotropic variants are themselves very close, suggesting that most of the gain from the full estimator comes from modeling off-diagonal covariance structure.

The fusion ablations show a similarly large effect. Using a posterior mean in place of the log-evidence combination reduces the score from $0.807$ to $0.708$, while removing prior anchoring gives $0.706$. In contrast, the logistic-head variants all fall between $0.801$ and $0.804$. Thus, the combination rule is substantially more consequential when applied to the Gaussian evidence than when applied to the logistic head.

| Variant | $k$=1 | 2 | 4 | 8 | 16 | 32 | 64 |
|---|---|---|---|---|---|---|---|
| **Ours** | **0.479** | **0.545** | 0.603 | 0.663 | 0.696 | 0.728 | 0.741 |
| free $\alpha_I, \alpha_T$ | 0.468 | 0.540 | **0.604** | **0.666** | **0.697** | **0.729** | **0.743** |
| $\alpha = 1$ | 0.479 | 0.545 | 0.603 | 0.663 | 0.696 | 0.728 | 0.741 |
| without prior anchoring | 0.479 | 0.545 | 0.603 | 0.663 | 0.696 | 0.728 | 0.741 |
| posterior mean | 0.466 | 0.533 | 0.593 | 0.654 | 0.688 | 0.718 | 0.735 |
| diagonal covariance | 0.471 | 0.535 | 0.593 | 0.652 | 0.682 | 0.703 | 0.709 |
| isotropic covariance | 0.471 | 0.534 | 0.592 | 0.651 | 0.681 | 0.702 | 0.708 |
| without the $(N-1)$ factor | 0.475 | 0.538 | 0.594 | 0.653 | 0.683 | 0.705 | 0.712 |
| logistic + shared $\alpha$ | 0.460 | 0.520 | 0.590 | 0.628 | 0.664 | 0.708 | 0.733 |
| logistic + $\alpha = 1$ | 0.460 | 0.520 | 0.590 | 0.628 | 0.664 | 0.708 | 0.733 |
| logistic + posterior mean | 0.460 | 0.520 | 0.589 | 0.627 | 0.663 | 0.707 | 0.733 |
| early fusion, closed form | 0.467 | 0.527 | 0.585 | 0.648 | 0.689 | 0.725 | 0.741 |
| text only (Gaussian) | 0.438 | 0.509 | 0.573 | 0.632 | 0.663 | 0.693 | 0.704 |
| image only (Gaussian) | 0.398 | 0.444 | 0.496 | 0.539 | 0.569 | 0.599 | 0.616 |
| text only (logistic) | 0.440 | 0.505 | 0.560 | 0.622 | 0.660 | 0.688 | 0.703 |
| image only (logistic) | 0.403 | 0.450 | 0.497 | 0.540 | 0.571 | 0.596 | 0.617 |

Table 11: Few-shot ablations, averaged over eight independently pretrained encoder pairs, eleven tasks, and three support-set resampling seeds.

### D.2 Few-Shot Training

Because the support set is balanced by construction, the empirical class prior is uniform. The shared coefficient and prior anchoring therefore do not change the predicted class, which produces identical results for FAME, $\alpha = 1$, and the variant without prior anchoring. Allowing separate modality coefficients changes performance only slightly across the shot counts.

The choice of per-modality classifier matters more. The Gaussian formulation outperforms the corresponding logistic head at every shot count, with the largest difference at $k = 8$ ($0.663$ versus $0.628$). The gap decreases as the labeling budget grows and is $0.008$ at $k = 64$. Full covariance also consistently outperforms the diagonal and isotropic variants, with the difference becoming larger at higher shot counts.

### D.3 Missingness in Test Data

FAME maintains the highest absolute score at every missingness rate. Freeing the modality-specific coefficients produces nearly identical results, while the posterior-mean and no-prior variants remain consistently lower. The same ordering is visible in the covariance ablations, with the full estimator outperforming both diagonal and isotropic covariance at every rate.

The contrast with early fusion becomes larger as missingness increases. From $r = 0$ to $r = 0.75$, zero- and mean-imputed early fusion decrease by $0.167$ and $0.140$, respectively, compared with $0.050$ for FAME. Because FAME combines evidence additively, an unavailable modality can be omitted rather than replaced by an imputed representation.

### D.4 Few-Shot Training Using Missing Modalities

The differences between fusion rules narrow in the combined setting. Separate modality coefficients slightly outperform the shared coefficient at the nonzero missingness rates, while the covariance ablations preserve the same ordering as in the other regimes.

Across the ablations, the full-covariance Gaussian estimator is the most consistent design choice: it improves over diagonal and isotropic covariance in every regime. Its advantage over the logistic head is largest in the few-shot setting and decreases as the labeling budget grows. The shared coefficient is generally sufficient, although separate modality coefficients provide a small improvement when few-shot supervision and modality missingness are combined.

| Variant | $r$=0 | 0.25 | 0.5 | 0.75 | $\Delta$ |
|---|---|---|---|---|---|
| *Fusion and weighting* | | | | | |
| **Ours** (shared $\alpha$) | **0.807** | **0.791** | **0.775** | **0.757** | −0.050 |
| free $\alpha_I, \alpha_T$ | 0.806 | 0.791 | 0.774 | 0.757 | −0.049 |
| $\alpha = 1$ | 0.756 | 0.733 | 0.710 | 0.687 | −0.069 |
| posterior mean | 0.708 | 0.690 | 0.670 | 0.650 | −0.058 |
| without prior anchoring | 0.706 | 0.686 | 0.666 | 0.645 | −0.061 |
| *Gaussian estimator* | | | | | |
| diagonal covariance | 0.742 | 0.728 | 0.715 | 0.697 | −0.046 |
| isotropic covariance | 0.739 | 0.724 | 0.711 | 0.693 | −0.046 |
| without the $(N-1)$ factor | 0.744 | 0.728 | 0.715 | 0.697 | −0.047 |
| *Alternative heads and combination rules* | | | | | |
| logistic + shared $\alpha$ | 0.802 | 0.785 | 0.767 | 0.748 | −0.054 |
| logistic + $\alpha = 1$ | 0.804 | 0.786 | 0.768 | 0.749 | −0.054 |
| logistic without prior anchoring | 0.801 | 0.781 | 0.761 | 0.741 | −0.060 |
| weighted posterior average | 0.726 | 0.705 | 0.686 | 0.665 | −0.061 |
| max rule | 0.705 | 0.691 | 0.676 | 0.661 | −0.044 |
| min rule | 0.682 | 0.641 | 0.599 | 0.559 | −0.123 |
| *Controls* | | | | | |
| early fusion, zero imputation | 0.782 | 0.723 | 0.667 | 0.615 | −0.167 |
| early fusion, mean imputation | 0.782 | 0.735 | 0.687 | 0.642 | −0.140 |
| ASIF | 0.675 | 0.628 | 0.593 | 0.558 | −0.117 |
| text only | 0.691 | 0.645 | 0.597 | 0.551 | −0.140 |
| image only | 0.629 | 0.590 | 0.549 | 0.508 | −0.121 |

Table 12: Ablations under missingness in test data, averaged over eight independently pretrained encoder pairs, eleven tasks, and three masking seeds. $\Delta$ denotes the change from $r = 0$ to $r = 0.75$.

# E Derivations and Properties of FaME

We derive the Gaussian class scores and prior-corrected fusion rule used by FaME, then explain the shared fusion coefficient and missing-modality prediction. These are standard Gaussian and product-rule identities (Wang et al., 2024; Kittler et al., 1998) applied to our fitted model; they do not assume that the fitted distributions equal the data-generating distributions.

## E.1 Gaussian discriminant scores

We give the GDA derivation used by Wang et al. (2024) in our notation, retaining a general class prior. Let $\mathcal{M} = \{1, \dots, M\}$ index modalities and $z^{(m)} \in \mathbb{R}^{d_m}$ denote their frozen representations. Condition on the fitted means $\widehat{\mu}_{m,c}$, positive class priors $\widehat{\pi}_c$ summing to one, and symmetric positive-definite precisions $\widehat{\Omega}_m$. Define

$$\widetilde{\Sigma}_m = \widehat{\Omega}_m^{-1}, \qquad g_{m,c}(z) = \mathcal{N}(z; \widehat{\mu}_{m,c}, \widetilde{\Sigma}_m). \tag{9}$$

Here $\widetilde{\Sigma}_m$ is the covariance implied by the regularized precision, distinct from the empirical covariance $\widehat{\Sigma}_m$. The trace-ridge estimator in the method is positive definite whenever $N_m > 1$, $\widehat{\Sigma}_m \succeq 0$, and $\mathrm{tr}(\widehat{\Sigma}_m) > 0$. Degenerate estimates require the fallback described in the implementation details.

**From class-conditional densities to posteriors.** For a fixed modality, Bayes' rule gives

$$\widehat{p}_m(c \mid z) = \frac{\widehat{\pi}_c g_{m,c}(z)}{\sum_k \widehat{\pi}_k g_{m,k}(z)} = \frac{\exp\{\log \widehat{\pi}_c + \log g_{m,c}(z)\}}{\sum_k \exp\{\log \widehat{\pi}_k + \log g_{m,k}(z)\}}. \tag{10}$$

Thus the posterior is a softmax of the log likelihood plus the log prior. The Gaussian density is

$$g_{m,c}(z) = \frac{\exp\Big\{-\frac{1}{2}(z - \widehat{\mu}_{m,c})^\top \widehat{\Omega}_m (z - \widehat{\mu}_{m,c})\Big\}}{(2\pi)^{d_m/2} |\widetilde{\Sigma}_m|^{1/2}}. \tag{11}$$

| Variant | $r$=0 | 0.25 | 0.5 | 0.75 | $\Delta$ |
|---|---|---|---|---|---|
| *Fusion and weighting* | | | | | |
| Ours (shared $\alpha$) | 0.650 | 0.641 | 0.601 | 0.590 | −0.060 |
| free $\alpha_I, \alpha_T$ | **0.651** | **0.646** | 0.606 | 0.599 | −0.052 |
| $\alpha = 1$ | 0.650 | 0.645 | 0.606 | 0.594 | −0.055 |
| posterior mean | 0.640 | 0.636 | 0.597 | 0.597 | −0.044 |
| without prior anchoring | 0.650 | 0.641 | 0.601 | 0.590 | −0.060 |
| *Gaussian estimator* | | | | | |
| diagonal covariance | 0.637 | 0.630 | 0.591 | 0.584 | −0.054 |
| isotropic covariance | 0.636 | 0.629 | 0.590 | 0.583 | −0.053 |
| without the $(N-1)$ factor | 0.639 | 0.632 | 0.593 | 0.585 | −0.053 |
| *Alternative heads and combination rules* | | | | | |
| logistic + shared $\alpha$ | 0.627 | 0.631 | 0.566 | 0.542 | −0.084 |
| logistic + $\alpha = 1$ | 0.627 | 0.632 | 0.548 | 0.560 | −0.066 |
| logistic without prior anchoring | 0.627 | 0.620 | 0.574 | 0.565 | −0.061 |
| weighted posterior average | 0.649 | 0.642 | 0.598 | 0.597 | −0.052 |
| max rule | 0.640 | 0.635 | 0.596 | 0.595 | −0.045 |
| min rule | 0.637 | 0.631 | 0.590 | 0.578 | −0.059 |
| *Controls* | | | | | |
| text only | 0.618 | 0.612 | 0.562 | 0.563 | −0.055 |
| image only | 0.532 | 0.520 | 0.491 | 0.491 | −0.041 |

Table 13: Few-shot training using missing modalities, averaged over eight independently pretrained encoder pairs and eleven tasks, pooled over $k \in \{4, 16\}$ and three masking seeds. ASIF is omitted because the few-shot pool does not provide the required number of complete image–text anchor pairs. $\Delta$ is defined as in Table 12.

Since $\widehat{\Omega}_m$ is symmetric, its quadratic form expands as

$$(z - \widehat{\mu}_{m,c})^\top \widehat{\Omega}_m (z - \widehat{\mu}_{m,c}) = z^\top \widehat{\Omega}_m z - 2 z^\top \widehat{\Omega}_m \widehat{\mu}_{m,c} + \widehat{\mu}_{m,c}^\top \widehat{\Omega}_m \widehat{\mu}_{m,c}. \tag{12}$$

Taking logarithms and substituting this expansion yields

$$\begin{aligned} \log g_{m,c}(z) &= -\frac{d_m}{2} \log(2\pi) - \frac{1}{2} \log\det \widetilde{\Sigma}_m - \frac{1}{2}(z - \widehat{\mu}_{m,c})^\top \widehat{\Omega}_m (z - \widehat{\mu}_{m,c}) \\ &= a_m(z) + \underbrace{z^\top \widehat{\Omega}_m \widehat{\mu}_{m,c} - \frac{1}{2} \widehat{\mu}_{m,c}^\top \widehat{\Omega}_m \widehat{\mu}_{m,c}}_{\widehat{\delta}_{m,c}(z)}, \end{aligned} \tag{13}$$

where $a_m(z) = -\frac{d_m}{2} \log(2\pi) - \frac{1}{2} \log\det \widetilde{\Sigma}_m - \frac{1}{2} z^\top \widehat{\Omega}_m z$ is independent of the class. The factor $\exp\{a_m(z)\}$ therefore cancels between the numerator and denominator:

$$\begin{aligned} \widehat{p}_m(c \mid z) &= \frac{\widehat{\pi}_c \exp\{a_m(z) + \widehat{\delta}_{m,c}(z)\}}{\sum_k \widehat{\pi}_k \exp\{a_m(z) + \widehat{\delta}_{m,k}(z)\}} \\ &= \frac{\exp\{\log \widehat{\pi}_c + \widehat{\delta}_{m,c}(z)\}}{\sum_k \exp\{\log \widehat{\pi}_k + \widehat{\delta}_{m,k}(z)\}}. \end{aligned} \tag{14}$$

The cancellation relies on covariance being shared across classes within each modality; covariances and representation dimensions may differ between modalities.

**Equivalent linear classifier.** Define

$$\widehat{w}_{m,c} = \widehat{\Omega}_m \widehat{\mu}_{m,c}, \qquad \widehat{b}_{m,c} = -\frac{1}{2} \widehat{\mu}_{m,c}^\top \widehat{\Omega}_m \widehat{\mu}_{m,c}. \tag{15}$$

Then $\widehat{\delta}_{m,c}(z) = z^\top \widehat{w}_{m,c} + \widehat{b}_{m,c}$, so the Gaussian posterior is a linear softmax classifier with weight $\widehat{w}_{m,c}$ and intercept $\widehat{b}_{m,c} + \log \widehat{\pi}_c$. These parameters are computed from the support statistics without gradient-based fitting. We keep the log prior separate from $\widehat{\delta}_{m,c}$ so it can be removed before fusion and included once in the combined score.

### E.2 PRIOR CORRECTION AND EXACT FUSION

Write $z = (z^{(1)}, \dots, z^{(M)})$ and define $g_m(u) = \sum_k \widehat{\pi}_k g_{m,k}(u)$. The evidence term is

$$\widehat{e}_{m,c}(u) = \log \widehat{p}_m(c \mid u) - \log \widehat{\pi}_c = \log g_{m,c}(u) - \log g_m(u). \tag{16}$$

**Proposition E.1** (Prior-corrected product rule)**.** *Under the fitted conditionally independent model $q(c, z) = \widehat{\pi}_c \prod_m g_{m,c}(z^{(m)})$, the posterior is*

$$q(c \mid z) = \frac{\exp\{\log \widehat{\pi}_c + \sum_m \widehat{e}_{m,c}(z^{(m)})\}}{\sum_k \exp\{\log \widehat{\pi}_k + \sum_m \widehat{e}_{m,k}(z^{(m)})\}}. \tag{17}$$

*Proof.* Bayes' rule and the definition of evidence give

$$q(c \mid z) \propto \widehat{\pi}_c \prod_m g_{m,c}(z^{(m)}) \propto \widehat{\pi}_c \prod_m \frac{\widehat{p}_m(c \mid z^{(m)})}{\widehat{\pi}_c}.$$

The omitted factor $\prod_m g_m(z^{(m)})$ is independent of $c$. Taking logarithms and normalizing proves the claim. □

Multiplying the modality posteriors without correction counts the prior $M$ times; dividing each by $\widehat{\pi}_c$ and restoring one prior gives the stated model. If the fitted priors and likelihoods equal the true ones and modalities are conditionally independent given the class, this posterior is the true posterior, and its maximizer is Bayes-optimal for zero-one loss. Otherwise, exactness applies only to the fitted model; the rule does not model class-dependent cross-modal interactions.

### E.3 EFFECT OF THE SHARED FUSION COEFFICIENT

For $\alpha > 0$, let $S_{\alpha,c}(z) = \log \widehat{\pi}_c + \alpha \sum_m \widehat{e}_{m,c}(z^{(m)})$ and let $q_\alpha(c \mid z)$ be its softmax over classes.

**Proposition E.2** (Equivalent Gaussian fusion rule)**.** *The fused probabilities satisfy*

$$q_\alpha(c \mid z) \propto \widehat{\pi}_c \prod_m g_{m,c}(z^{(m)})^\alpha \propto \exp\left\{ \log \widehat{\pi}_c + \alpha \sum_m \widehat{\delta}_{m,c}(z^{(m)}) \right\}. \tag{18}$$

*They are the posterior of a factorized Gaussian model with the same means and prior and covariances $\widetilde{\Sigma}_m/\alpha$.*

*Proof.* Substitute the evidence identity into $S_{\alpha,c}$. The terms $-\alpha \sum_m \log g_m(z^{(m)})$ are independent of $c$, as are the terms $\alpha \sum_m a_m(z^{(m)})$ in the Gaussian expansion. Dropping them gives the two equivalent expressions. Moreover,

$$g_{m,c}(u)^\alpha = K_m(\alpha)\, \mathcal{N}(u; \widehat{\mu}_{m,c}, \widetilde{\Sigma}_m/\alpha),$$

where $K_m(\alpha)$ is independent of $c$ because the covariance is class-shared. These factors cancel on normalization, proving the Gaussian interpretation. □

Thus the prediction can be evaluated directly from the discriminants, without computing modality posteriors. The coefficient changes the strength of the combined evidence relative to the prior, but does not change the relative weights of modalities. With uniform priors, $\arg\max_c S_{\alpha,c}(z) = \arg\max_c \sum_m \widehat{\delta}_{m,c}(z^{(m)})$ for every $\alpha > 0$: class predictions are unchanged under fixed tie-breaking, although probabilities can change. The covariance identity does not itself guarantee calibration to the true posterior.

### E.4 MISSING MODALITIES

**Proposition E.3** (Marginalizing unavailable modalities)**.** *For a fixed subset $A \subseteq \mathcal{M}$ of available modalities, the fitted model has posterior*

$$q(c \mid z_A) \propto \widehat{\pi}_c \prod_{m \in A} g_{m,c}(z^{(m)}) \propto \exp\left\{ \log \widehat{\pi}_c + \sum_{m \in A} \widehat{e}_{m,c}(z^{(m)}) \right\}. \tag{19}$$

*For general $\alpha > 0$, weighting the available evidence by $\alpha$ gives the corresponding posterior under the rescaled Gaussian model above.*

*Proof.* Integrating $q(c, z)$ over each unavailable representation removes its density because $\int g_{m,c}(u)\, du = 1$. Applying Bayes' rule and the evidence identity gives the result. The same argument applies to the normalized Gaussian densities with covariances $\widetilde{\Sigma}_m/\alpha$. □

This justifies omitting missing evidence for a fixed subset. If the observation pattern is random, ignoring it in the true posterior further requires that it provide no class information conditional on the observed features.

At adaptation time, each modality's means and covariance use only its own observed labeled representations, so complete pairs are unnecessary. For these observations to represent the target class-conditional distribution, a sufficient condition is that modality availability be independent of its feature values given the class, with positive observation probability in every class. Independent fitting alone does not correct selection bias from informative missingness.

# F Additional Results

## F.1 Full Data Results

We report complete results for the full data setting, where the complete training data is used, in Table 14. Results are averaged over the eight encoder pairs whose image and text encoders were pretrained independently; the jointly pretrained CLIP-CLIP pair is considered separately in Appendix F.2. For each dataset, we report its primary metric: accuracy for single label tasks and macro one-vs-rest average precision for multilabel tasks. Results are averaged first across tasks within each dataset and then across encoder pairs. Methods trained with gradient descent are averaged over three seeds.

| Method | CUB | COCO | VG | CrisisMMD | MM-IMDb | MIMIC | Mean |
|---|---|---|---|---|---|---|---|
| *No gradient based training* | | | | | | | |
| Late fusion (mean) | 0.860 | 0.805 | 0.527 | 0.797 | 0.624 | 0.710 | 0.720 |
| ASIF | 0.464 | 0.907 | 0.538 | 0.969 | 0.382 | 0.368 | 0.604 |
| **Ours** | **0.861** | 0.933 | 0.656 | 0.986 | 0.625 | 0.710 | **0.795** |
| *Trained fusion head* | | | | | | | |
| Early fusion (lin) | 0.840 | 0.927 | 0.656 | 0.984 | 0.508 | 0.523 | 0.740 |
| Early fusion (MLP) | 0.830 | 0.940 | 0.669 | 0.988 | 0.604 | 0.648 | 0.780 |
| GMU | 0.841 | 0.941 | 0.667 | 0.989 | 0.565 | 0.640 | 0.774 |
| TFN | 0.808 | 0.933 | 0.647 | 0.988 | 0.632 | 0.726 | 0.789 |
| LMF | 0.831 | 0.933 | 0.639 | 0.987 | 0.614 | 0.735 | 0.790 |
| *Encoder adaptation* | | | | | | | |
| LoRA | 0.842 | 0.942 | 0.681 | 0.993 | 0.661 | 0.866 | 0.831 |
| Full fine-tune | 0.835 | 0.932 | 0.681 | 0.993 | 0.667 | **0.910** | **0.836** |

Table 14: Results with the full training set, averaged over the eight independently pretrained encoder pairs. Methods trained with gradient descent are additionally averaged over three seeds. Bold denotes the best overall result and the best result among methods without gradient based training. The jointly pretrained CLIP-CLIP pair is reported separately in Appendix F.2.

With the full training set, FAME achieves a mean score of $0.795$, compared with $0.720$ for mean late fusion and $0.604$ for ASIF. It also exceeds the trained fusion heads on average, whose mean scores range from $0.740$ to $0.790$. Encoder adaptation unsurprisingly performs best overall, reaching $0.831$ with LoRA and $0.836$ with full fine-tuning. This suggests that when multimodal paired data is available, there is power in training end-to-end.

**Unimodal classifiers and linear probes.** We next examine how much information is available from each modality before fusion. The *Image only* and *Text only* baselines use the same Gaussian

discriminant as FAME, but make predictions from only the image representation or only the text representation, respectively. We compare these with discriminative linear probes trained on the same frozen representations. These controls separate the quality of the classifier within each modality from the benefit of combining evidence across modalities.

| Method | CUB | COCO | VG | CrisisMMD | MM-IMDb | MIMIC | Mean |
|---|---|---|---|---|---|---|---|
| Image only (Gaussian) | 0.856 | 0.777 | 0.509 | 0.741 | 0.401 | 0.287 | 0.595 |
| Linear probe (image) | 0.839 | 0.911 | 0.601 | 0.851 | 0.415 | 0.266 | 0.647 |
| Text only (Gaussian) | 0.552 | 0.813 | 0.527 | 0.917 | 0.586 | 0.713 | 0.685 |
| Linear probe (text) | 0.351 | 0.929 | 0.633 | 0.987 | 0.586 | 0.666 | 0.692 |
| **Ours** | 0.861 | 0.933 | 0.656 | 0.986 | 0.625 | 0.710 | **0.795** |

Table 15: Unimodal Gaussian classifiers and linear probes in the full data setting, averaged over eight independently pretrained encoder pairs. Image only and Text only use the same Gaussian discriminant as FAME but receive only the indicated modality. Linear probes are trained discriminatively on the same frozen representations. FAME combines the image and text Gaussian evidence.

Table 15 shows that the performance of FAME cannot be explained by uniformly stronger classifiers within each modality. On average, the image linear probe outperforms the Gaussian image classifier ($0.647$ versus $0.595$), while the two text classifiers perform similarly ($0.692$ versus $0.685$). Their relative performance also varies substantially across datasets. For example, the Gaussian text classifier performs better on CUB ($0.552$ versus $0.351$), whereas the linear probe performs better on COCO and CrisisMMD.

Despite these differences within individual modalities, combining the two Gaussian evidence terms yields a mean score of $0.795$, substantially above either unimodal classifier. The full data results therefore indicate that the main benefit of FAME comes from combining complementary class evidence across modalities rather than from a consistent advantage of the Gaussian classifier within either modality.

### F.2 Results with Jointly Pretrained CLIP Encoders

Our main experiments consider image and text encoders that were pretrained independently. One pair in the encoder grid is different: the CLIP image and text encoders were trained jointly with a contrastive objective and therefore share a common representation space. We exclude this pair from the main averages so that the primary evaluation reflects the setting targeted by FAME, where modalities may come from unrelated pretrained encoders. For completeness, we report results for the CLIP-CLIP pair separately. This provides a robustness check for the case in which the two modality representations were learned jointly. We evaluate the pair under the same full data, few shot, and missingness settings used in the main experiments.

Under full supervision, FAME achieves a mean score of $0.777$, compared with $0.684$ for mean late fusion and $0.569$ for ASIF. Its performance is also comparable to the trained fusion heads, whose means range from $0.726$ to $0.775$. Encoder adaptation performs best overall, reaching $0.807$ with LoRA and $0.800$ with full fine-tuning. As in the independently pretrained encoder results, much of the difference between FAME and encoder adaptation comes from MIMIC-CXR. Excluding MIMIC-CXR, FAME, LoRA, and full fine-tuning obtain mean scores of $0.812$, $0.821$, and $0.809$, respectively. On MIMIC-CXR itself, FAME reaches $0.604$, while LoRA and full fine-tuning reach $0.737$ and $0.754$.

The same pattern holds when supervision is limited. FAME performs best at each evaluated shot count, increasing from $0.396$ at $k = 1$ to $0.715$ at $k = 64$. Mean late fusion is the closest baseline across these budgets, while the trained fusion heads are generally less effective in the lowest data regimes.

Under test time missingness, FAME retains the highest absolute score at every masking rate, decreasing from $0.777$ at $r = 0$ to $0.714$ at $r = 0.75$. The few shot missingness results show a similar pattern, with FAME remaining competitive across masking rates. These results indicate that the additive evidence formulation remains effective even when the two representations come from jointly pretrained encoders.

| Method | CUB | COCO | VG | CrisisMMD | MM-IMDb | MIMIC | Mean |
|---|---|---|---|---|---|---|---|
| *No gradient based training* | | | | | | | |
| Image only | 0.825 | 0.751 | 0.513 | 0.741 | 0.483 | 0.273 | 0.598 |
| Text only | 0.561 | 0.786 | 0.509 | 0.891 | 0.557 | 0.601 | 0.651 |
| Late fusion (mean) | 0.833 | 0.768 | 0.520 | 0.742 | 0.634 | 0.604 | 0.684 |
| ASIF | 0.417 | 0.905 | 0.562 | 0.945 | 0.305 | 0.277 | 0.569 |
| **Ours** | 0.830 | 0.935 | 0.679 | 0.980 | 0.634 | 0.604 | **0.777** |
| *Trained fusion head* | | | | | | | |
| Early fusion (lin) | 0.789 | 0.924 | 0.681 | 0.985 | 0.533 | 0.446 | 0.726 |
| Early fusion (MLP) | 0.800 | 0.941 | 0.688 | 0.987 | 0.619 | 0.549 | 0.764 |
| GMU | 0.805 | 0.941 | 0.685 | 0.989 | 0.581 | 0.535 | 0.756 |
| TFN | 0.760 | 0.942 | 0.682 | 0.988 | 0.644 | 0.611 | 0.771 |
| LMF | 0.782 | 0.940 | 0.680 | 0.986 | 0.641 | 0.620 | 0.775 |
| *Encoder adaptation* | | | | | | | |
| LoRA | 0.781 | 0.950 | 0.701 | 0.996 | 0.676 | 0.737 | **0.807** |
| Full fine-tune | 0.766 | 0.934 | 0.690 | 0.995 | 0.662 | 0.754 | 0.800 |
| *Unimodal linear probes* | | | | | | | |
| Probe (image) | 0.795 | 0.925 | 0.648 | 0.901 | 0.497 | 0.255 | 0.670 |
| Probe (text) | 0.288 | 0.932 | 0.642 | 0.984 | 0.566 | 0.569 | 0.664 |

Table 16: Full data results for the jointly pretrained CLIP image and text encoders. Values are averaged over tasks within each dataset. Because this evaluation contains a single encoder pair, there is no averaging across encoder pairs. Bold marks the best mean among methods without gradient based training and the best mean overall.

| Method | $k$=1 | $k$=4 | $k$=16 | $k$=64 |
|---|---|---|---|---|
| Image only | 0.320 | 0.454 | 0.549 | 0.596 |
| Text only | 0.355 | 0.503 | 0.606 | 0.651 |
| Late fusion (mean) | 0.380 | 0.545 | 0.654 | 0.709 |
| Early fusion (lin) | 0.332 | 0.430 | 0.543 | 0.614 |
| Early fusion (MLP) | 0.242 | 0.445 | 0.559 | 0.582 |
| GMU | 0.341 | 0.468 | 0.564 | 0.587 |
| TFN | 0.244 | 0.373 | 0.503 | 0.595 |
| LMF | 0.234 | 0.424 | 0.509 | 0.571 |
| **Ours** | **0.396** | **0.559** | **0.666** | **0.715** |

Table 17: Few shot results for CLIP-CLIP, averaged over three resampling seeds. ASIF is omitted because it requires at least $50$ complete image-text anchor pairs.

**Comparison with independently pretrained encoders.** The CLIP-CLIP results are broadly consistent with those obtained from the eight independently pretrained encoder pairs, but they should not be interpreted as an isolated test of representation alignment. Both the pretraining procedure and the underlying encoders change between these comparisons.

With full data, FAME obtains $0.777$ on CLIP-CLIP compared with $0.795$ when averaged over the independently pretrained pairs. Its few shot scores are also modestly lower than the corresponding averages from the main evaluation. These differences do not indicate that joint pretraining itself is harmful; they may instead reflect differences in the quality and downstream suitability of the individual encoders. We similarly do not observe a systematic advantage for methods that can make use of correspondence between the two modalities. ASIF obtains a mean of $0.569$ on CLIP-CLIP, while TFN and LMF obtain $0.771$ and $0.775$, respectively. Overall, the conclusions of the main evaluation continue to hold for the jointly pretrained pair, while this comparison does not isolate the effect of shared representation geometry from differences in the underlying encoder representations.

### F.3 Results Breakdown by Encoder Pair

The main results average performance across eight pairs of independently pretrained image and text encoders. Here, we report each encoder pair separately to assess whether the aggregate results are

| Method | $r$=0 | 0.25 | 0.5 | 0.75 | $\Delta$ |
|---|---|---|---|---|---|
| *Test time missingness* | | | | | |
| Image only | 0.598 | 0.560 | 0.521 | 0.481 | −0.117 |
| Text only | 0.651 | 0.608 | 0.560 | 0.515 | −0.136 |
| Late fusion (mean) | 0.684 | 0.665 | 0.645 | 0.624 | −0.059 |
| Early fusion (lin) | 0.726 | 0.669 | 0.646 | 0.631 | −0.096 |
| Early fusion (MLP) | 0.764 | 0.721 | 0.685 | 0.651 | −0.113 |
| GMU | 0.756 | 0.717 | 0.680 | 0.642 | −0.114 |
| TFN | 0.771 | 0.721 | 0.678 | 0.633 | −0.138 |
| LMF | 0.775 | 0.724 | 0.676 | 0.628 | −0.147 |
| ASIF | 0.569 | 0.530 | 0.504 | 0.482 | −0.086 |
| **Ours** | **0.777** | **0.757** | **0.736** | **0.714** | **−0.063** |
| *Few shot setting with missingness* | | | | | |
| Image only | 0.498 | 0.489 | 0.442 | 0.422 | −0.076 |
| Text only | 0.539 | 0.529 | 0.493 | 0.486 | −0.053 |
| Late fusion (mean) | 0.590 | 0.582 | 0.544 | 0.535 | −0.055 |
| Early fusion (lin) | 0.461 | 0.432 | 0.362 | 0.359 | −0.102 |
| Early fusion (MLP) | 0.476 | 0.458 | 0.422 | 0.389 | −0.087 |
| GMU | 0.496 | 0.481 | 0.450 | 0.399 | −0.097 |
| TFN | 0.417 | 0.415 | 0.331 | 0.304 | −0.113 |
| LMF | 0.451 | 0.435 | 0.364 | 0.379 | −0.072 |
| **Ours** | **0.601** | **0.591** | **0.553** | **0.537** | −0.064 |

Table 18: Missingness results for CLIP-CLIP, averaged over three masking seeds. $\Delta$ denotes the change from $r = 0$ to $r = 0.75$. ASIF is omitted from the few shot setting because the support pool does not provide the required number of complete image-text anchor pairs.

consistent across representation choices. The jointly pretrained CLIP-CLIP pair is excluded from these tables and reported separately in Appendix F.2. Each entry reports the mean and standard deviation over three seeds. For unimodal baselines, repeated values across pairs sharing the same image or text encoder are expected because the unused modality does not affect prediction.

**Few-Shot Training** We construct few-shot training sets with $k \in \{1, 2, 4, 8, 16, 32, 64\}$ labeled examples per class and evaluate on the full test set. The validation set is subsampled to the same labeling budget so that model selection does not use additional labeled supervision. We report the encoder-pair results at $k = 4$ and $k = 16$ in Tables 19 and 20, respectively.

Across encoder choices, FAME performs best for all eight pairs at both $k = 4$ and $k = 16$. The mean score across encoder pairs is $0.574$ at $k = 4$ and $0.682$ at $k = 16$, compared with $0.500$ and $0.606$ for mean late fusion. The few-shot improvement is therefore not driven by a particular image or text representation.

**Missingness in Test Data** Training data retain both modalities, while validation and test examples are independently missing one modality with probability $r \in \{0, 0.25, 0.5, 0.75\}$. When an example is masked, either its image or text modality is removed, but never both. We apply the same missingness rate to the validation and test sets so that model selection is performed under the same observation conditions encountered at inference time. We report the encoder-pair results at $r = 0.5$ and $r = 0.75$ in Tables 21 and 22, respectively.

The encoder-level results are particularly consistent under missingness in test data. FAME performs best for all eight encoder pairs at both reported missingness rates. Averaged across encoder pairs, it reaches $0.746$ at $r = 0.5$ and $0.720$ at $r = 0.75$, compared with $0.619$ and $0.594$ for mean late fusion. The improvement therefore persists across image and text representation choices.

**Few-Shot Training Using Missing Modalities** We apply the missingness procedure from the previous setting to few-shot training sets with $k \in \{4, 16\}$. This setting evaluates performance when both labeled data and complete multimodal observations are limited. We report the encoder-pair results at $r = 0.5$ and $r = 0.75$ in Tables 23 and 24, respectively, pooling over $k \in \{4, 16\}$.

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.454 \pm 0.012$ | $0.454 \pm 0.012$ |
| Text only | $0.528 \pm 0.022$ | $0.515 \pm 0.030$ |
| Late fusion (mean) | $0.496 \pm 0.023$ | $0.492 \pm 0.029$ |
| Early fusion (lin) | $0.406 \pm 0.014$ | $0.425 \pm 0.024$ |
| Early fusion (MLP) | $0.343 \pm 0.054$ | $0.391 \pm 0.029$ |
| GMU | $0.443 \pm 0.022$ | $0.439 \pm 0.043$ |
| TFN | $0.345 \pm 0.053$ | $0.403 \pm 0.027$ |
| LMF | $0.431 \pm 0.013$ | $0.466 \pm 0.030$ |
| **Ours** | $\mathbf{0.581 \pm 0.023}$ | $\mathbf{0.575 \pm 0.028}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.443 \pm 0.015$ | $0.443 \pm 0.015$ | $0.443 \pm 0.015$ |
| Text only | $0.503 \pm 0.015$ | $0.528 \pm 0.022$ | $0.515 \pm 0.030$ |
| Late fusion (mean) | $0.505 \pm 0.019$ | $0.519 \pm 0.023$ | $0.513 \pm 0.027$ |
| Early fusion (lin) | $0.450 \pm 0.014$ | $0.444 \pm 0.018$ | $0.445 \pm 0.015$ |
| Early fusion (MLP) | $0.424 \pm 0.008$ | $0.423 \pm 0.038$ | $0.453 \pm 0.042$ |
| GMU | $0.477 \pm 0.005$ | $0.447 \pm 0.007$ | $0.446 \pm 0.025$ |
| TFN | $0.386 \pm 0.062$ | $0.414 \pm 0.015$ | $0.437 \pm 0.032$ |
| LMF | $0.481 \pm 0.023$ | $0.435 \pm 0.033$ | $0.476 \pm 0.023$ |
| **Ours** | $\mathbf{0.565 \pm 0.025}$ | $\mathbf{0.591 \pm 0.027}$ | $\mathbf{0.588 \pm 0.030}$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.429 \pm 0.009$ | $0.429 \pm 0.009$ | $0.429 \pm 0.009$ |
| Text only | $0.503 \pm 0.015$ | $0.528 \pm 0.022$ | $0.515 \pm 0.030$ |
| Late fusion (mean) | $0.484 \pm 0.017$ | $0.498 \pm 0.022$ | $0.497 \pm 0.025$ |
| Early fusion (lin) | $0.424 \pm 0.011$ | $0.419 \pm 0.009$ | $0.431 \pm 0.038$ |
| Early fusion (MLP) | $0.392 \pm 0.015$ | $0.394 \pm 0.040$ | $0.439 \pm 0.033$ |
| GMU | $0.446 \pm 0.022$ | $0.439 \pm 0.007$ | $0.450 \pm 0.010$ |
| TFN | $0.365 \pm 0.062$ | $0.413 \pm 0.012$ | $0.431 \pm 0.004$ |
| LMF | $0.471 \pm 0.008$ | $0.429 \pm 0.025$ | $0.477 \pm 0.010$ |
| **Ours** | $\mathbf{0.547 \pm 0.021}$ | $\mathbf{0.575 \pm 0.025}$ | $\mathbf{0.573 \pm 0.026}$ |

Table 19: Few-shot results at $k = 4$, reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three support set resampling seeds. Bold denotes the best mean for each encoder pair.

The combined setting is more sensitive to the representation pair. At $r = 0.5$, FAME has the highest mean score across encoder pairs ($0.527$ versus $0.517$ for mean late fusion) and performs best on seven of the eight pairs. At $r = 0.75$, the two methods are closer: mean late fusion averages $0.522$ and FAME averages $0.518$, with each performing best on four encoder pairs. Thus, the strongest consistency appears in the few-shot and test-missingness settings, while performance under simultaneous data and modality scarcity depends more strongly on the underlying representations.

## G IMAGE ATTRIBUTION

Images from Figure 1 were obtained from https://commons.wikimedia.org/wiki/Main_Page. We have included images from the following users:

- https://commons.wikimedia.org/wiki/User:Rhododendrites
- https://commons.wikimedia.org/wiki/User:Diliff
- https://commons.wikimedia.org/wiki/User:Tisha_Mukherjee
- https://commons.wikimedia.org/wiki/User:Zcebeci

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.549 \pm 0.006$ | $0.549 \pm 0.006$ |
| Text only | $0.637 \pm 0.002$ | $0.627 \pm 0.002$ |
| Late fusion (mean) | $0.616 \pm 0.003$ | $0.611 \pm 0.001$ |
| Early fusion (lin) | $0.555 \pm 0.005$ | $0.558 \pm 0.010$ |
| Early fusion (MLP) | $0.534 \pm 0.016$ | $0.539 \pm 0.001$ |
| GMU | $0.537 \pm 0.016$ | $0.554 \pm 0.025$ |
| TFN | $0.435 \pm 0.050$ | $0.497 \pm 0.053$ |
| LMF | $0.504 \pm 0.025$ | $0.545 \pm 0.004$ |
| **Ours** | $\mathbf{0.698 \pm 0.002}$ | $\mathbf{0.693 \pm 0.001}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.501 \pm 0.006$ | $0.501 \pm 0.006$ | $0.501 \pm 0.006$ |
| Text only | $0.606 \pm 0.003$ | $0.637 \pm 0.002$ | $0.627 \pm 0.002$ |
| Late fusion (mean) | $0.596 \pm 0.002$ | $0.622 \pm 0.002$ | $0.617 \pm 0.001$ |
| Early fusion (lin) | $0.552 \pm 0.008$ | $0.552 \pm 0.008$ | $0.567 \pm 0.009$ |
| Early fusion (MLP) | $0.546 \pm 0.013$ | $0.544 \pm 0.022$ | $0.565 \pm 0.001$ |
| GMU | $0.566 \pm 0.003$ | $0.522 \pm 0.017$ | $0.567 \pm 0.003$ |
| TFN | $0.529 \pm 0.020$ | $0.464 \pm 0.019$ | $0.533 \pm 0.012$ |
| LMF | $0.543 \pm 0.010$ | $0.488 \pm 0.023$ | $0.544 \pm 0.012$ |
| **Ours** | $\mathbf{0.660 \pm 0.003}$ | $\mathbf{0.699 \pm 0.003}$ | $\mathbf{0.695 \pm 0.001}$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.479 \pm 0.006$ | $0.479 \pm 0.006$ | $0.479 \pm 0.006$ |
| Text only | $0.606 \pm 0.003$ | $0.637 \pm 0.002$ | $0.627 \pm 0.002$ |
| Late fusion (mean) | $0.580 \pm 0.001$ | $0.605 \pm 0.002$ | $0.600 \pm 0.001$ |
| Early fusion (lin) | $0.545 \pm 0.007$ | $0.538 \pm 0.007$ | $0.557 \pm 0.015$ |
| Early fusion (MLP) | $0.530 \pm 0.018$ | $0.524 \pm 0.011$ | $0.552 \pm 0.002$ |
| GMU | $0.538 \pm 0.018$ | $0.517 \pm 0.003$ | $0.550 \pm 0.006$ |
| TFN | $0.514 \pm 0.010$ | $0.444 \pm 0.033$ | $0.512 \pm 0.011$ |
| LMF | $0.526 \pm 0.012$ | $0.481 \pm 0.011$ | $0.531 \pm 0.012$ |
| **Ours** | $\mathbf{0.645 \pm 0.002}$ | $\mathbf{0.686 \pm 0.003}$ | $\mathbf{0.681 \pm 0.002}$ |

Table 20: Few-shot results at $k = 16$, reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three support set resampling seeds. Bold denotes the best mean for each encoder pair.

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.521 \pm 0.004$ | $0.521 \pm 0.004$ |
| Text only | $0.604 \pm 0.002$ | $0.588 \pm 0.001$ |
| Late fusion (mean) | $0.640 \pm 0.002$ | $0.621 \pm 0.002$ |
| Early fusion (lin) | $0.624 \pm 0.002$ | $0.657 \pm 0.003$ |
| Early fusion (MLP) | $0.678 \pm 0.003$ | $0.698 \pm 0.002$ |
| GMU | $0.683 \pm 0.004$ | $0.697 \pm 0.002$ |
| TFN | $0.670 \pm 0.006$ | $0.689 \pm 0.003$ |
| LMF | $0.676 \pm 0.004$ | $0.694 \pm 0.002$ |
| **Ours** | $\mathbf{0.764 \pm 0.002}$ | $\mathbf{0.753 \pm 0.001}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.529 \pm 0.004$ | $0.529 \pm 0.004$ | $0.529 \pm 0.004$ |
| Text only | $0.560 \pm 0.001$ | $0.604 \pm 0.002$ | $0.588 \pm 0.001$ |
| Late fusion (mean) | $0.596 \pm 0.001$ | $0.645 \pm 0.002$ | $0.626 \pm 0.002$ |
| Early fusion (lin) | $0.638 \pm 0.001$ | $0.621 \pm 0.001$ | $0.664 \pm 0.002$ |
| Early fusion (MLP) | $0.676 \pm 0.003$ | $0.674 \pm 0.004$ | $0.695 \pm 0.003$ |
| GMU | $0.678 \pm 0.001$ | $0.686 \pm 0.003$ | $0.700 \pm 0.003$ |
| TFN | $0.676 \pm 0.003$ | $0.668 \pm 0.009$ | $0.699 \pm 0.002$ |
| LMF | $0.682 \pm 0.005$ | $0.665 \pm 0.008$ | $0.703 \pm 0.003$ |
| **Ours** | $\mathbf{0.729 \pm 0.001}$ | $\mathbf{0.762 \pm 0.001}$ | $\mathbf{0.753 \pm 0.001}$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.507 \pm 0.004$ | $0.507 \pm 0.004$ | $0.507 \pm 0.004$ |
| Text only | $0.560 \pm 0.001$ | $0.604 \pm 0.002$ | $0.588 \pm 0.001$ |
| Late fusion (mean) | $0.582 \pm 0.001$ | $0.633 \pm 0.001$ | $0.614 \pm 0.001$ |
| Early fusion (lin) | $0.632 \pm 0.001$ | $0.608 \pm 0.002$ | $0.650 \pm 0.002$ |
| Early fusion (MLP) | $0.670 \pm 0.003$ | $0.667 \pm 0.004$ | $0.688 \pm 0.002$ |
| GMU | $0.668 \pm 0.002$ | $0.674 \pm 0.003$ | $0.688 \pm 0.002$ |
| TFN | $0.667 \pm 0.001$ | $0.669 \pm 0.007$ | $0.687 \pm 0.001$ |
| LMF | $0.673 \pm 0.001$ | $0.680 \pm 0.006$ | $0.690 \pm 0.002$ |
| **Ours** | $\mathbf{0.717 \pm 0.000}$ | $\mathbf{0.751 \pm 0.001}$ | $\mathbf{0.741 \pm 0.001}$ |

Table 21: Missingness in test data at $r = 0.5$, reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three masking seeds. Bold denotes the best mean for each encoder pair.

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.481 \pm 0.001$ | $0.481 \pm 0.001$ |
| Text only | $0.551 \pm 0.004$ | $0.537 \pm 0.003$ |
| Late fusion (mean) | $0.612 \pm 0.003$ | $0.596 \pm 0.002$ |
| Early fusion (lin) | $0.605 \pm 0.000$ | $0.637 \pm 0.001$ |
| Early fusion (MLP) | $0.637 \pm 0.001$ | $0.660 \pm 0.002$ |
| GMU | $0.639 \pm 0.002$ | $0.653 \pm 0.003$ |
| TFN | $0.612 \pm 0.002$ | $0.636 \pm 0.002$ |
| LMF | $0.619 \pm 0.006$ | $0.643 \pm 0.004$ |
| **Ours** | $\mathbf{0.738 \pm 0.002}$ | $\mathbf{0.729 \pm 0.001}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.488 \pm 0.001$ | $0.488 \pm 0.001$ | $0.488 \pm 0.001$ |
| Text only | $0.515 \pm 0.003$ | $0.551 \pm 0.004$ | $0.537 \pm 0.003$ |
| Late fusion (mean) | $0.574 \pm 0.002$ | $0.616 \pm 0.003$ | $0.601 \pm 0.002$ |
| Early fusion (lin) | $0.615 \pm 0.002$ | $0.592 \pm 0.009$ | $0.637 \pm 0.001$ |
| Early fusion (MLP) | $0.637 \pm 0.001$ | $0.629 \pm 0.002$ | $0.653 \pm 0.001$ |
| GMU | $0.637 \pm 0.001$ | $0.640 \pm 0.001$ | $0.654 \pm 0.001$ |
| TFN | $0.627 \pm 0.002$ | $0.608 \pm 0.002$ | $0.644 \pm 0.003$ |
| LMF | $0.634 \pm 0.002$ | $0.603 \pm 0.002$ | $0.650 \pm 0.002$ |
| **Ours** | $\mathbf{0.705 \pm 0.002}$ | $\mathbf{0.734 \pm 0.003}$ | $\mathbf{0.726 \pm 0.002}$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.470 \pm 0.002$ | $0.470 \pm 0.002$ | $0.470 \pm 0.002$ |
| Text only | $0.515 \pm 0.003$ | $0.551 \pm 0.004$ | $0.537 \pm 0.003$ |
| Late fusion (mean) | $0.560 \pm 0.001$ | $0.603 \pm 0.002$ | $0.587 \pm 0.002$ |
| Early fusion (lin) | $0.608 \pm 0.003$ | $0.585 \pm 0.001$ | $0.625 \pm 0.001$ |
| Early fusion (MLP) | $0.633 \pm 0.001$ | $0.623 \pm 0.001$ | $0.644 \pm 0.002$ |
| GMU | $0.629 \pm 0.001$ | $0.630 \pm 0.003$ | $0.644 \pm 0.002$ |
| TFN | $0.623 \pm 0.002$ | $0.618 \pm 0.001$ | $0.634 \pm 0.001$ |
| LMF | $0.627 \pm 0.001$ | $0.622 \pm 0.001$ | $0.637 \pm 0.001$ |
| **Ours** | $\mathbf{0.691 \pm 0.001}$ | $\mathbf{0.722 \pm 0.002}$ | $\mathbf{0.712 \pm 0.001}$ |

Table 22: Missingness in test data at $r = 0.75$, reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three masking seeds. Bold denotes the best mean for each encoder pair.

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.403 \pm 0.031$ | $0.392 \pm 0.027$ |
| Text only | $0.449 \pm 0.055$ | $0.415 \pm 0.053$ |
| Late fusion (mean) | $0.488 \pm 0.040$ | $0.465 \pm 0.036$ |
| Early fusion (lin) | $0.246 \pm 0.004$ | $0.317 \pm 0.033$ |
| Early fusion (MLP) | $0.230 \pm 0.025$ | $0.290 \pm 0.027$ |
| GMU | $0.280 \pm 0.032$ | $0.358 \pm 0.048$ |
| TFN | $0.259 \pm 0.048$ | $0.312 \pm 0.012$ |
| LMF | $0.325 \pm 0.048$ | $0.382 \pm 0.021$ |
| **Ours** | $\mathbf{0.514 \pm 0.042}$ | $\mathbf{0.484 \pm 0.037}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.410 \pm 0.029$ | $0.406 \pm 0.017$ | $0.408 \pm 0.030$ |
| Text only | $0.435 \pm 0.035$ | $0.487 \pm 0.009$ | $0.415 \pm 0.053$ |
| Late fusion (mean) | $\mathbf{0.509 \pm 0.029}$ | $0.553 \pm 0.005$ | $0.502 \pm 0.034$ |
| Early fusion (lin) | $0.365 \pm 0.040$ | $0.283 \pm 0.039$ | $0.360 \pm 0.023$ |
| Early fusion (MLP) | $0.351 \pm 0.013$ | $0.330 \pm 0.052$ | $0.407 \pm 0.046$ |
| GMU | $0.391 \pm 0.039$ | $0.312 \pm 0.026$ | $0.387 \pm 0.042$ |
| TFN | $0.324 \pm 0.027$ | $0.362 \pm 0.032$ | $0.401 \pm 0.033$ |
| LMF | $0.415 \pm 0.027$ | $0.367 \pm 0.030$ | $0.422 \pm 0.049$ |
| **Ours** | $0.505 \pm 0.046$ | $\mathbf{0.562 \pm 0.003}$ | $\mathbf{0.504 \pm 0.044}$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.394 \pm 0.015$ | $0.421 \pm 0.020$ | $0.396 \pm 0.015$ |
| Text only | $0.467 \pm 0.005$ | $0.475 \pm 0.054$ | $0.513 \pm 0.008$ |
| Late fusion (mean) | $0.532 \pm 0.007$ | $0.526 \pm 0.036$ | $0.561 \pm 0.006$ |
| Early fusion (lin) | $0.394 \pm 0.020$ | $0.309 \pm 0.018$ | $0.405 \pm 0.026$ |
| Early fusion (MLP) | $0.405 \pm 0.014$ | $0.384 \pm 0.027$ | $0.431 \pm 0.012$ |
| GMU | $0.418 \pm 0.022$ | $0.325 \pm 0.004$ | $0.417 \pm 0.029$ |
| TFN | $0.352 \pm 0.054$ | $0.364 \pm 0.030$ | $0.401 \pm 0.018$ |
| LMF | $0.426 \pm 0.016$ | $0.381 \pm 0.040$ | $0.434 \pm 0.013$ |
| **Ours** | $\mathbf{0.533 \pm 0.004}$ | $\mathbf{0.534 \pm 0.045}$ | $\mathbf{0.577 \pm 0.001}$ |

Table 23: Few-shot training using missing modalities at $r = 0.5$, pooled over $k \in \{4, 16\}$ and reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three seeds. Bold denotes the best mean for each encoder pair.

| *CLIP image encoder* | | |
|---|---|---|
| Method | E5 | MPNet |
| Image only | $0.407 \pm 0.036$ | $0.398 \pm 0.037$ |
| Text only | $0.459 \pm 0.039$ | $0.435 \pm 0.027$ |
| Late fusion (mean) | $0.501 \pm 0.022$ | $0.480 \pm 0.019$ |
| Early fusion (lin) | $0.253 \pm 0.041$ | $0.328 \pm 0.038$ |
| Early fusion (MLP) | $0.210 \pm 0.006$ | $0.252 \pm 0.015$ |
| GMU | $0.253 \pm 0.058$ | $0.332 \pm 0.035$ |
| TFN | $0.284 \pm 0.054$ | $0.303 \pm 0.064$ |
| LMF | $0.271 \pm 0.054$ | $0.339 \pm 0.039$ |
| **Ours** | $\mathbf{0.512 \pm 0.018}$ | $\mathbf{0.484 \pm 0.023}$ |

| *DINOv2 image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.412 \pm 0.037$ | $0.398 \pm 0.031$ | $0.410 \pm 0.036$ |
| Text only | $0.440 \pm 0.040$ | $0.480 \pm 0.008$ | $0.435 \pm 0.027$ |
| Late fusion (mean) | $\mathbf{0.528 \pm 0.014}$ | $0.543 \pm 0.009$ | $\mathbf{0.512 \pm 0.018}$ |
| Early fusion (lin) | $0.380 \pm 0.030$ | $0.300 \pm 0.024$ | $0.383 \pm 0.048$ |
| Early fusion (MLP) | $0.364 \pm 0.041$ | $0.378 \pm 0.013$ | $0.365 \pm 0.055$ |
| GMU | $0.405 \pm 0.048$ | $0.277 \pm 0.077$ | $0.333 \pm 0.026$ |
| TFN | $0.322 \pm 0.040$ | $0.324 \pm 0.036$ | $0.403 \pm 0.009$ |
| LMF | $0.410 \pm 0.046$ | $0.376 \pm 0.026$ | $0.403 \pm 0.040$ |
| **Ours** | $0.503 \pm 0.031$ | $\mathbf{0.547 \pm 0.003}$ | $0.503 \pm 0.026$ |

| *ViT image encoder* | | | |
|---|---|---|---|
| Method | CLIP | E5 | MPNet |
| Image only | $0.378 \pm 0.022$ | $0.433 \pm 0.025$ | $0.381 \pm 0.022$ |
| Text only | $0.458 \pm 0.002$ | $0.489 \pm 0.042$ | $0.505 \pm 0.009$ |
| Late fusion (mean) | $0.419 \pm 0.007$ | $0.440 \pm 0.022$ | $0.450 \pm 0.008$ |
| Early fusion (lin) | $0.408 \pm 0.013$ | $0.337 \pm 0.042$ | $0.395 \pm 0.024$ |
| Early fusion (MLP) | $0.368 \pm 0.023$ | $0.353 \pm 0.020$ | $0.422 \pm 0.036$ |
| GMU | $0.384 \pm 0.012$ | $0.307 \pm 0.011$ | $0.358 \pm 0.068$ |
| TFN | $0.355 \pm 0.041$ | $0.343 \pm 0.035$ | $0.371 \pm 0.030$ |
| LMF | $0.435 \pm 0.015$ | $0.392 \pm 0.047$ | $0.433 \pm 0.006$ |
| **Ours** | $\mathbf{0.501 \pm 0.019}$ | $\mathbf{0.534 \pm 0.035}$ | $\mathbf{0.557 \pm 0.002}$ |

Table 24: Few-shot training using missing modalities at $r = 0.75$, pooled over $k \in \{4, 16\}$ and reported separately for each image and text encoder pair. Each entry is mean $\pm$ standard deviation over three seeds. Bold denotes the best mean for each encoder pair.