Do More Modalities Always Help? A Geometric Perspective on Missing-Modality Robustness

Songyuan Sui[1] **Zhen Tan**[2] **Mohan Zhang**[3]
Rana Muhammad Shahroz Khan[3] **Xia Hu**[1] **Tianlong Chen**[3]
[1]Rice University [2]Stevens Institute of Technology
[3]University of North Carolina at Chapel Hill
Songyuan.Sui@rice.edu

Abstract

Missing modality remains a longstanding challenge in multimodal learning. Existing methods typically address this issue through modality recovery or adaptive strategies. However, they overlook models' internal cross-modal dependencies formed during multimodal training, which later impair robustness. We systematically characterize a counterintuitive deployment-time failure mode: models trained on full modalities can underperform unimodal models when one modality is missing at inference time. This pattern appears across diverse architectures, such as fusion models, CLIP-style two-tower models, and vision-language models. We show that such degradation is closely associated with learned cross-modal dependencies in the principal parameter subspaces. Multimodal training induces structured rotations of these subspaces, particularly in cross-modal interaction layers. These rotations are associated with reduced task-aligned margins and larger task-aware representation harm under missing-modality inputs. We propose **Geodesic Unlearning (GU)**, a lightweight parameter-editing method that leverages Grassmannian subspace geometry for structured subspace correction to improve missing-modality robustness. It rotates the principal input subspace toward a unimodal reference along a geodesic path. We prove that this correction minimizes the distance to the reference within a fixed subspace-distance budget. Experiments across architectures and datasets show that GU improves performance under missing-modality inference while preserving full-modality accuracy, outperforming strong missing-modality robustness baselines. These findings support a geometric view of deployment-time missing-modality degradation and suggest localized subspace editing as a practical route for robustness correction.

1 Introduction

Multimodal learning has achieved strong performance across language, vision, and healthcare domains [66, 8, 44]. Large language models (LLMs) also support reasoning and interpretation beyond plain text, such as vision, time series, and relational databases [8, 46, 18, 45]. In jointly trained multimodal models, combining complementary signals typically improves performance over their unimodal counterparts [21]. However, this advantage may not transfer to deployment, where a modality can become unavailable due to sensor failure, incomplete data collection, privacy constraints, or system limitations [58]. Prior work on imbalanced multimodal learning shows that joint training may under-optimize modality-specific representations [37]. We study a distinct deployment-time manifestation: when one modality is completely absent at inference, a model trained with full modalities can underperform one trained only on the available modality. This setting usually arises in heterogeneous deployment environments with varying device or sensor availability [35, 40, 13].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Across diverse architectures, we observe a recurring ordering where full-modality inference performs best, unimodal training performs next, and full-modality training with a missing modality performs worst. This suggests that missing-modality degradation is not merely due to absent input information, but is also linked to internal cross-modal dependencies shaped by multimodal training.

We characterize this degradation geometrically. At the parameter level, multimodal models exhibit structured principal-subspace rotations, measured by principal-angle deviations between matched multimodal and unimodal input subspaces, especially in cross-modal interaction layers. Rotation magnitude is strongly associated with the missing-modality gap. At the representation level, misclassified samples exhibit larger task-aware harm (negative signed task margin) than correctly classified samples at intermediate multimodal interaction stages under missing-modality inference. This gap persists toward the classifier. These localized parameter- and representation-space signatures motivate targeted parameter-level correction beyond recovery-based or adaptive methods.

Motivated by these findings, we formulate missing-modality robustness as a special form of *machine unlearning*: attenuating harmful cross-modal dependencies under missing inputs while preserving useful multimodal functionality. We propose **Geodesic Unlearning (GU)**, a lightweight post-training editing method. Given a multimodal model and a matched unimodal reference, GU edits **one target layer** selected by an angle–harm criterion. It treats the selected principal subspace as a point on the *Grassmannian*, rotates its input-side basis toward the unimodal reference along a geodesic path, and reconstructs the layer while retaining the principal coefficients, output factor, and residual. Our theoretical analysis establishes the geometric optimality of this correction under a fixed subspace-distance budget and shows that the path advances all principal angles proportionally.

Experiments across fusion models, CLIP-style two-tower architectures, vision-language models (VLMs), and multiple datasets further show that GU improves performance on missing-modality tasks while largely preserving full-modality accuracy. Comparisons with multimodal-specific methods, general training and adaptation baselines, and parameter-editing controls support GU's favorable robustness–preservation trade-off, with the strongest gains when harmful dependencies are localized in editable interaction layers.

In summary, our contributions are as follows:

- We systematically characterize deployment-time degradation when one modality is entirely unavailable at inference across three major multimodal architectures.
- We provide a geometric account of this degradation, supported by parameter- and representation-level evidence, showing that it is strongly associated with principal subspace rotations and task-aware representational harm in cross-modal interaction layers.
- We propose Geodesic Unlearning, a lightweight, theoretically grounded Grassmannian subspace editing method that applies across models to improve missing-modality robustness while preserving full-modality utility.

2 Related work

Missing Modality remains a challenge in multimodal learning [43, 68, 58]. Existing approaches address this challenge by recovering unavailable information or improving how models use available modalities. Recovery-based methods impute missing inputs at the data level [63, 22, 49, 54, 69] or representation level [35, 53, 74, 47, 12, 51], relying on cross-modal correlations and reconstruction quality. Adaptive methods accommodate missing inputs through flexible architectures [61, 50, 65], parameter adaptation [19, 73], or training–inference strategies [52, 64, 20, 60, 57, 9]. Complementary modality-balancing methods, including OGM-GE [37], AGM [30], DnR [56], and MCR [27], regulate the relative learning contributions of modalities during joint training. Our work focuses on harmful parameter dependencies retained when a modality is completely absent at deployment. We diagnose their task-relevant principal-subspace geometry and correct them in an already-trained model, providing a model-side approach that can complement training-time robustness objectives.

Geometric Understanding of Neural Networks has been studied from both representation and parameter perspectives. At the representation level, prior work reveals structured feature geometry, including simplex class structure [36], representational similarity across models [29], and margin-based geometry related to generalization and robustness [11, 34]. At the parameter level, recent studies show that model parameters and gradients often concentrate in low-dimensional subspaces, motivating

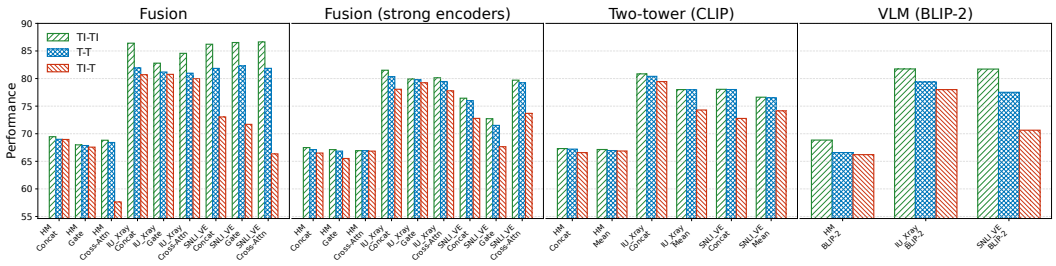


Figure 1: Missing-modality degradation across architectures and datasets. F1 scores are reported for TI-TI, T-T, and TI-T settings. In 26 of the 27 settings, we observe TI-TI $>$ T-T $>$ TI-T, revealing a recurring deployment-time missing-modality degradation pattern.

low-rank adaptation and subspace-based optimization [19, 72, 41]. More recently, Grassmannian methods model the evolution of such subspaces and enable structure-preserving updates through geodesic paths [10, 17, 39]. Our work connects these geometric perspectives to missing-modality robustness by showing that multimodal training induces structured rotations in principal parameter subspaces, which are associated with task-aware representation harm under missing inputs.

Machine Unlearning (MU) was originally proposed in the privacy domain to remove the influence of specific training data without full retraining [4, 3]. Existing methods span data-side interventions [67, 48], training-time updates [5, 25], decoding-time control [14, 38], and training-free parameter editing [24, 2]. Beyond privacy, MU has been explored to improve model behavior in scenarios such as continual learning [55], domain adaptation [1], and domain generalization [70]. Motivated by our finding that multimodal training induces harmful cross-modal dependencies under missing inputs, we repurpose MU to selectively attenuate these dependencies while preserving useful multimodal capability. To the best of our knowledge, our work is the first to formulate missing-modality robustness as an unlearning problem and address it through geometric parameter editing.

3 Problem setup and preliminaries

3.1 Deployment-time missing modality setting

We consider a multimodal learning problem over modality set $\mathcal{M} = \{1, \ldots, M\}$, where each modality $m \in \mathcal{M}$ has input space $\mathcal{X}_m$. The full-modality input space is $\mathcal{X}_{\text{full}} = \prod_{m \in \mathcal{M}} \mathcal{X}_m$, and a full input is denoted by $x = (x_1, \ldots, x_M)$. We study the **deployment-time** missing modality setting, where models are trained on full-modality inputs but evaluated with one or more modalities unavailable at inference. We introduce a null symbol $\perp$ and define $\mathcal{X}_{\text{miss}} = \prod_{m \in \mathcal{M}} (\mathcal{X}_m \cup \{\perp\})$. If modality m is missing at deployment, the corresponding input is $x_{-m} = (x_1, \ldots, x_{m-1}, \perp, x_{m+1}, \ldots, x_M)$. In practice, $\perp$ can be implemented as a masking token, zeroed feature stream, or learned null embedding. We focus on the practically important **single missing modality** case. For bimodal systems such as text-image models, the same modality is **entirely** absent for every test example in the missing-modality evaluation, while the other remains available.

3.2 Evaluation protocol and missing-modality gap

We instantiate the setting with bimodal text-image models, where **T** denotes text and **I** denotes image. We consider three training–testing regimes: **TI-TI** trains and tests with text+image; **T-T** trains and tests with text only; and **TI-T** trains with text+image but tests with text only. Unless otherwise stated, we study missing-image evaluation, where text remains available, and the image is removed at inference. We define the **missing-modality gap** as $\Delta_{\text{miss}} = \text{Perf}(\text{TI-TI}) - \text{Perf}(\text{TI-T})$ and the **unimodal robustness gap** as $\Delta_{\text{uni}} = \text{Perf}(\text{T-T}) - \text{Perf}(\text{TI-T})$, where $\text{Perf}(\cdot)$ denotes the task metric, instantiated as F1 scores in our main experiments. The former measures performance loss after removing the image modality at deployment, while the latter measures the extent to which the missing-modality multimodal model underperforms a text-specialized model.

Table 1: Representative model families used in our analysis. We cover distinct multimodal integration mechanisms, including explicit fusion architectures, aligned two-tower encoders, and VLMs.

Model Family	Backbone(s)	Variants
Fusion	RoBERTa [33] + ResNet50 [28]	Concat, Gate, Cross-Attention
Fusion (strong encoders)	Qwen [62] + OpenCLIP [23]	Concat, Gate, Cross-Attention
Two-tower	OpenCLIP	Mean Pooling, Late Concat
VLM	BLIP-2 [31]	Classifier Head

3.3 Parameter subspace geometry

To study how multimodal training changes layer parameters, we analyze the principal singular subspaces of selected weight matrices. For a layer weight matrix $W \in \mathbb{R}^{d_o \times d_i}$ with SVD $W = U\Sigma V^\top$, we use the top-k singular vectors $U_k = [u_1, \ldots, u_k]$ and $V_k = [v_1, \ldots, v_k]$ to define the principal output and input subspaces, respectively. The right singular vectors V_k identify the dominant input directions the layer reads, while U_k specifies the associated output directions. Since V_k and $V_k Q$ span the same subspace for any orthogonal $Q \in \mathbb{R}^{k \times k}$, the basis-invariant object is $S = \mathrm{span}(V_k) \in \mathrm{Gr}(k, d_i)$; the output subspace is defined analogously on $\mathrm{Gr}(k, d_o)$. Given a multimodal model W_{TI} and a unimodal reference model W_{T}, we compare their corresponding input and output subspaces to quantify how multimodal training changes parameter geometry, using standard metrics such as principal angles, subspace overlap, and Frobenius distances.

4 Empirical analysis of missing-modality degradation

4.1 Analysis scope: architectures and benchmarks

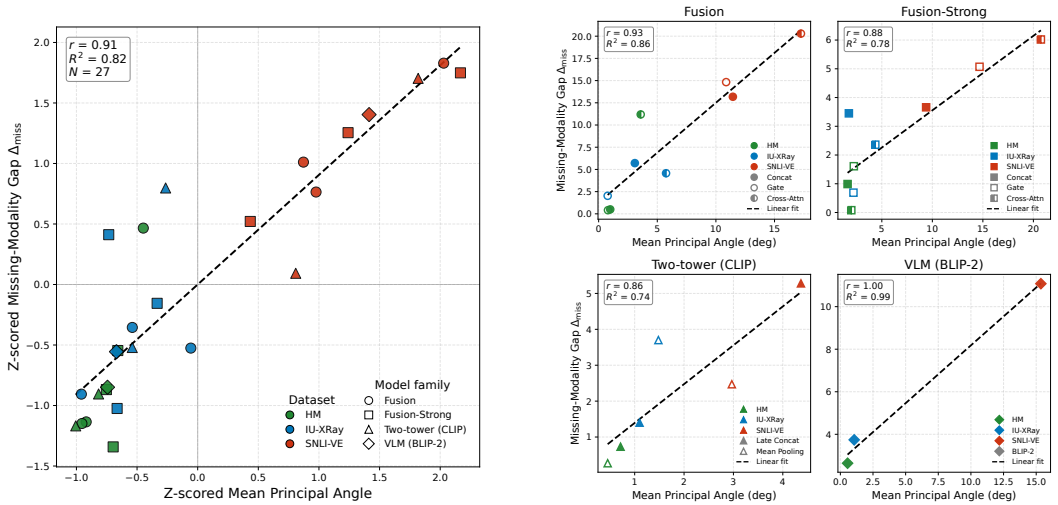
As summarized in Table 1, our analysis covers three representative families of multimodal architectures: fusion models, two-tower dual-encoder models, and vision-language models. This partition follows recent surveys that categorize multimodal systems by architectural organization and fusion strategy [71, 16, 32]. The selected systems differ in how they encode, align, and fuse modalities. It allows us to examine whether deployment-time missing-modality degradation is tied to a specific architecture or appears more broadly. All analyses are conducted under the deployment-time missing-image setting defined in Section 3.2. Dataset details, evaluation metrics, and implementation protocols are provided in Section 6.

4.2 RQ1: Is missing-modality degradation consistent across architectures?

Observation 1. Across architectures and datasets, full-modality inference performs best, while missing-modality inference generally underperforms unimodal training.

Prior work on imbalanced multimodal learning has shown that joint multimodal training may underoptimize modality-specific representations [37]. We examine a distinct deployment-time question: whether a model trained with full modalities remains robust when one modality is entirely unavailable at inference. Figure 1 compares the three regimes in Section 3.2 across 27 architecture–dataset combinations from Table 1. The ordering $\mathrm{Perf}(\text{TI-TI}) > \mathrm{Perf}(\text{T-T}) > \mathrm{Perf}(\text{TI-T})$ holds in 26 of 27 settings, while $\mathrm{Perf}(\text{T-T}) > \mathrm{Perf}(\text{TI-T})$ holds in all 27 settings. The complete numerical results are provided in Appendix C.1.

Across the 27 settings, the missing-modality gap averages 4.74 F1 points but ranges from 0.08 to 20.29. Generally, architectures with more explicit cross-modal interaction, such as cross-attention fusion, tend to exhibit larger missing-modality gaps than those with weaker cross-modal coupling, such as gated fusion or mean-pooled dual encoders. This suggests that degradation is not only a consequence of removing input information, but is also associated with internal cross-modal dependencies formed during multimodal training. These dependencies may be analyzed from a parameter-subspace perspective.



(a) Normalized correlation across model families.

(b) Raw correlation within each architecture.

Figure 2: Principal subspace rotation correlates with missing-modality degradation. (a) Within-family z-score normalization reveals a global positive association. (b) Raw within-family measurements show consistent positive associations without normalization.

4.3 RQ2: Does parameter geometry track degradation?

Observation 2. Larger principal subspace rotations in cross-modal interaction layers are strongly and positively associated with larger missing-modality gaps.

We examine whether the degradation in Section 4.2 has a measurable parameter-space signature. For each candidate layer, we compare the TI-trained model with its T-only reference using the principal singular subspaces defined in Section 3.3. We measure their divergence by the mean principal angle $\bar{\theta} = k^{-1} \sum_{i=1}^{k} \theta_i$, where $\theta_i = \arccos \sigma_i \left(V_{\mathrm{TI},k}^{\top} V_{\mathrm{T},k} \right)$ and $\sigma_i(\cdot)$ denotes the ith singular value.

Figure 2 shows that interaction-layer subspace rotation strongly tracks missing-modality degradation. After applying within-family z-score normalization to account for scale differences, Figure 2(a) shows a strong positive association between mean principal angle and missing-modality gap over all 27 architecture–dataset settings ($r = 0.91$, $R^2 = 0.82$). The same relationship also appears in raw within-family measurements, as shown in Figure 2(b). It includes fusion models ($r = 0.93$, $R^2 = 0.86$), stronger-encoder fusion models ($r = 0.88$, $R^2 = 0.78$), and CLIP-style two-tower models ($r = 0.86$, $R^2 = 0.74$). The BLIP-2 trend is also positive, but should be interpreted descriptively because it contains only three points. On SNLI-VE, Fusion Cross-Attention exhibits a gap of 20.29 F1 points, with mean input/output principal angles of $16.00°/16.57°$ at `attn_v_from_t_in_proj`, versus $1.95°/0.00°$ at the final classifier (Appendix C.2).

These results indicate a strong positive association that is not driven solely by normalization. They also suggest that multimodal training induces nonuniform parameter shifts, with the most informative geometric changes concentrated in cross-modal interaction layers. Principal-angle deviation therefore provides a geometric signature of missing-modality failure. The strength and consistency of this association motivate interaction-layer subspaces as localized targets for robustness correction.

4.4 RQ3: Where does representation harm concentrate?

Observation 3. Final-error versus final-correct task-aware harm is consistently represented at intermediate stages of multimodal interaction and often persists near the classifier.

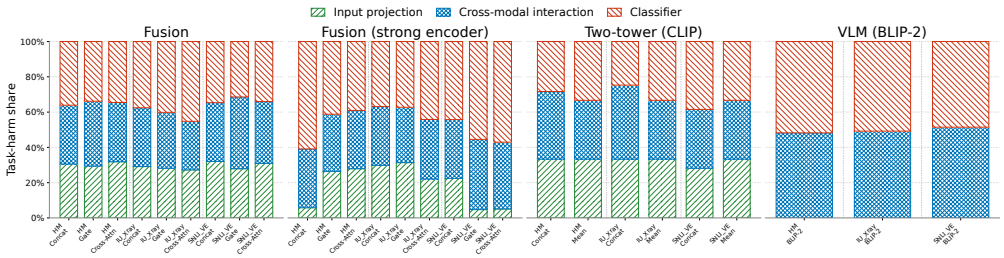


Figure 3: Stage-wise distribution of task-aware representation harm. The final-error versus final-correct harm gap appears at cross-modal interaction stages and often persists near the classifier.

We next localize task-relevant representation shifts induced by missing modality inputs. For our binary classification tasks, we use a centroid-margin diagnostic at each probed stage. Let μ_+ and μ_- be the class centroids computed from representations at that stage, with task direction $d = \mu_+ - \mu_-$ and midpoint $c = (\mu_+ + \mu_-)/2$. For representation $h(x)$ and label $y \in \{+1, -1\}$, we define the signed task margin as $m(x) = y \cdot (h(x) - c)^\top \hat{d}$, where $\hat{d} = d/\|d\|_2$, and define $\mathrm{Harm}(x) = -m(x)$. The error and correct sets, $\mathcal{E}$ and $\mathcal{C}$, are defined by the final prediction under missing-modality evaluation. We then probe intermediate representations using the same partition and compute $\Delta_{\mathrm{harm}} = \mathbb{E}_{x\in\mathcal{E}}[\mathrm{Harm}(x)] - \mathbb{E}_{x\in\mathcal{C}}[\mathrm{Harm}(x)]$.

Figure 3 summarizes the stage-wise distribution of the final-error versus final-correct harm gap. Bars are normalized within each dataset and model setting, so they indicate where the gap concentrates rather than its absolute magnitude. Final-error samples exhibit larger harm near the classifier, but the gap is already evident at cross-modal interaction stages before the final decision layer. These stages account for a substantial share of the gap across settings, consistent with Section 4.3, where the strongest subspace rotations also appear in cross-modal interaction layers. Classifier-stage harm in several settings is therefore better interpreted as downstream persistence than as a classifier-only effect. Detailed layer-wise results are provided in Appendix C.3.

Together, Sections 4.2–4.4 provide the empirical basis for our method design: degradation is broadly observed, strongly associated with interaction-layer subspace rotations, and accompanied by task-aware representation shifts concentrated around multimodal interaction stages. We therefore use the identified interaction-layer subspace as a localized target for robustness correction.

5 Geodesic Unlearning

5.1 Overview

Section 4 identifies cross-modal interaction layers as localized targets. Their principal subspace rotations track missing-modality degradation, and their representations exhibit task-aware harm. Given a trained multimodal model and a matched unimodal reference, we propose **Geodesic Unlearning (GU)** to perform a lightweight post-training correction of one selected layer. GU selects the target layer using an angle-harm criterion, rotates its principal input-side subspace toward the corresponding unimodal subspace along a Grassmannian geodesic, and reconstructs the layer while retaining the principal coefficients, output factor, and residual structure. This selectively attenuates harmful cross-modal dependencies under missing inputs while preserving useful multimodal functionality. The same subspace-editing procedure applies to fusion, CLIP-style two-tower, and BLIP-2 models, with only the target layer varying by architecture, enabling a unified post-training correction.

5.2 Target-layer selection

For each model, GU selects one target layer from the interaction and nearby projection/classifier layers analyzed in Section 4. For each candidate layer ℓ, let θ_ℓ denote the mean principal angle between the top-k input subspaces of the TI model and T reference, and let $\Delta_{\mathrm{harm},\ell}$ denote the final-error versus final-correct harm gap measured at layer ℓ under missing-modality evaluation. We rank layers on the development set using

$$s_\ell = z(\theta_\ell) + z(\Delta_{\mathrm{harm},\ell}), \tag{1}$$

where $z(\cdot)$ denotes z-score normalization across candidate layers within the same model. The layer with the largest s_ℓ is edited; the test set is reserved only for final evaluation.

5.3 Subspace realignment and reconstruction

Given the selected multimodal layer weight $W_{\mathrm{TI}} \in \mathbb{R}^{d_o \times d_i}$ with SVD $W_{\mathrm{TI}} = U_{\mathrm{TI}}\Sigma_{\mathrm{TI}}V_{\mathrm{TI}}^\top$, we decompose it into a top-k principal component and a residual:

$$W_{\mathrm{TI}} = U_{\mathrm{TI},k}\Sigma_{\mathrm{TI},k}V_{\mathrm{TI},k}^\top + R_{\mathrm{TI}}, \qquad R_{\mathrm{TI}} = W_{\mathrm{TI}} - U_{\mathrm{TI},k}\Sigma_{\mathrm{TI},k}V_{\mathrm{TI},k}^\top. \tag{2}$$

Our diagnostics localize task-relevant deviation in the principal input subspace, so GU edits this orientation. Let $V_{\mathrm{T},k}$ be an orthonormal basis of the corresponding top-k input subspace of the matched unimodal reference. GU computes an orthonormal basis for the edited subspace

$$V(\eta) = \mathrm{Geodesic}(V_{\mathrm{TI},k}, V_{\mathrm{T},k}; \eta), \qquad \eta \in [0,1], \tag{3}$$

where η controls the edit strength: $\eta = 0$ leaves the multimodal subspace unchanged, while $\eta = 1$ reaches the unimodal-reference endpoint along the geodesic path. The explicit geodesic construction is provided in Appendix D.1. The edited parameter is reconstructed as

$$W' = R_{\mathrm{TI}} + U_{\mathrm{TI},k}\Sigma_{\mathrm{TI},k}V(\eta)^\top. \tag{4}$$

The reconstruction retains the TI output factor $U_{\mathrm{TI},k}$, principal coefficients $\Sigma_{\mathrm{TI},k}$, and residual R_{TI}, whereas direct weight interpolation also changes these components.

Geometric rationale. Let $S_{\mathrm{TI}} = \mathrm{span}(V_{\mathrm{TI},k})$, $S_{\mathrm{T}} = \mathrm{span}(V_{\mathrm{T},k})$, and $S_\eta = \mathrm{span}(V(\eta))$. Write d for the Grassmann geodesic distance, the ℓ_2 norm of the principal-angle vector, and $D = d(S_{\mathrm{TI}}, S_{\mathrm{T}})$. For any candidate subspace S satisfying $d(S_{\mathrm{TI}}, S) \le \eta D$, the minimizing geodesic gives

$$d(S_{\mathrm{TI}}, S_\eta) = \eta D, \qquad d(S_\eta, S_{\mathrm{T}}) = (1-\eta)D \le d(S, S_{\mathrm{T}}). \tag{5}$$

Thus, GU reaches a subspace as close to the reference as any edit within the same subspace-distance budget. It advances through an η fraction of each principal angle, whereas linear interpolation of aligned bases followed by orthonormalization generally yields angle-dependent progress. Appendix D.2 provides the proof and the linear-path comparison. Controlled comparisons of target-layer selection and editing strategies are reported in Appendix F.5.

6 Experiments

6.1 Experimental setup

Datasets and metrics. We evaluate GU on three multimodal benchmarks: *Hateful Memes* [26] for social multimodal classification, *SNLI-VE* [59] for visual entailment, and *IU-XRay* [6] for medical vision-language prediction. These three benchmarks are formulated as binary classification tasks under the deployment-time missing-modality setting. We report F1 as the primary metric to account for potential class imbalance. More dataset information is provided in Appendix E.1. We further evaluate GU on *A-OKVQA* [42], a multiple-choice VQA benchmark, in Appendix F.6.

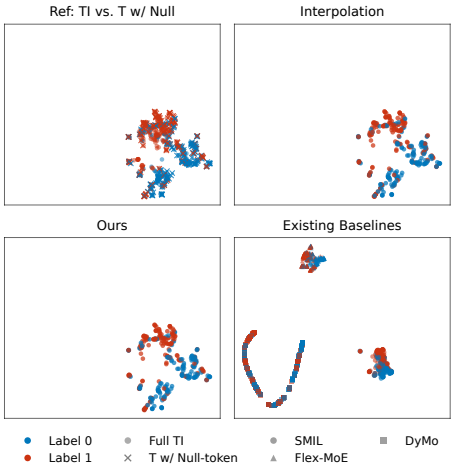
Model families. We use the model families summarized in Table 1. For each dataset and model family, we train a full multimodal model with text and image inputs, denoted as the TI model. We train a matched T reference using the same overall parameterization, where the image branch is replaced by a null input during both training and inference. This implements the T-T regime while keeping layer shapes aligned for subspace comparison and editing.

Baselines. Our comparisons cover three groups. **Multimodal-specific methods** include SMIL [35], Flex-MoE [65], and DyMo [9] in a unified fusion setting. We further evaluate MoRA [73] and OGM-GE [37] on compatible architectures beyond this setting. **General training and adaptation baselines** include post-hoc LoRA [19] and Dual CE. Dual CE jointly supervises full-input and missing-image paths. **Parameter-editing controls** include layer interpolation (Weight), following parameter-space editing ideas [24], and aligned-subspace linear interpolation (Linear). These controls isolate the effects of the edited object and interpolation path. Appendix E.2 details their definitions, architecture coverage, and protocols. Additional comparisons are reported in Appendices F.4 and F.5.

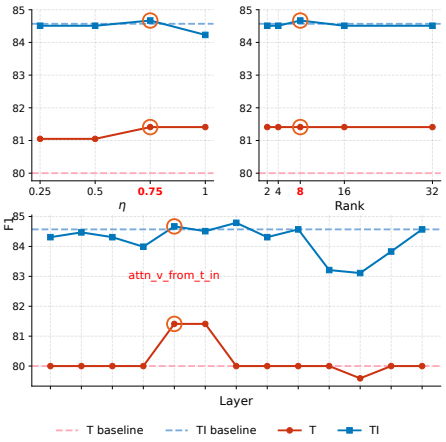
Implementation. GU is applied after training. For the main comparison with missing-modality baselines, we use a unified RoBERTa–ResNet50 cross-attention fusion setting. For cross-family

Table 2: Main missing-image performance (TI-T, F1 ↑). Values are three-run means. Underline and **bold** denote the second-best and best results, and parentheses show gains over the second best.

Dataset	Null-token	SMIL	Flex-MoE	DyMo	Interpolation	**GU (Ours)**
HM	57.63	56.60	47.58	48.82	58.50	**60.88** (+2.38)
IU-XRay	80.00	79.21	71.02	79.19	80.14	**81.41** (+1.27)
SNLI-VE	66.35	72.20	70.74	61.30	67.29	**74.41** (+2.21)



(a) UMAP of fusion-layer representations.



(b) Ablations of η, rank, and target layer.

Figure 4: Representation structure and ablation analysis. (a) On SNLI-VE, GU and layer interpolation preserve similar broad UMAP layouts. (b) On IU-XRay, target-layer selection has the largest effect on robustness recovery.

robustness and preservation evaluation, we apply GU and layer interpolation to all model families in Table 1. We evaluate missing-image recovery alongside full-modality preservation. Baseline protocols and implementation details are provided in Appendices E.2 and E.4.

6.2 Main results in the deployment-time missing-modality setting

Performance. Table 2 reports mean F1 over three independent runs under the TI-T setting, where the image modality is entirely absent at inference. Corresponding sample standard deviations and supplementary ten-seed paired significance tests are provided in Appendix F.1. GU achieves the highest mean F1 among the compared methods on all three datasets, improving over the strongest baseline in each dataset by +2.38 on Hateful Memes, +1.27 on IU-XRay, and +2.21 on SNLI-VE. These consistent gains are obtained by editing only one selected layer after training. Thus, GU's advantage comes from directly addressing the failure mode identified in Section 4: harmful cross-modal dependencies in localized parameter subspaces.

The results also highlight why this deployment setting is challenging for data-side recovery and adaptive fusion methods. Many missing-modality methods are designed for settings with more available modalities and partial sample-level missingness, where the remaining modalities can still provide sufficient cross-modal information for recovery, routing, or adaptive fusion. In our setting, the system is bimodal and one of the two modalities is completely unavailable, leaving only text information. Under this stricter condition, baselines are less consistently effective: SMIL and Flex-MoE are competitive on SNLI-VE but do not consistently improve over the null-token baseline across datasets, while DyMo degrades substantially. By contrast, GU works from the model side, correcting the parameter subspace associated with harmful multimodal dependencies rather than attempting to reconstruct an unavailable modality. This explains why a simple null-token pathway can remain competitive on Hateful Memes and IU-XRay, while GU further improves it through targeted parameter-space correction.

Table 3: Family-level robustness–preservation trade-off on SNLI-VE. ΔT and ΔTI denote changes from the original TI model in TI-T and TI-TI, respectively.

Family	#Var	TI-T		TI-TI		ΔT		ΔTI	
		Interp.	**GU**	Interp.	**GU**	Interp.	**GU**	Interp.	**GU**
Fusion	3	71.34	**73.86**	86.33	**86.37**	0.98	**3.49**	−0.13	**−0.09**
Fusion (str.)	3	72.34	**73.06**	75.16	**75.95**	0.96	**1.69**	−1.13	**−0.34**
CLIP	2	74.50	**74.68**	76.13	**76.70**	1.04	**1.22**	−1.21	**−0.64**
VLM	1	70.65	**70.65**	81.72	**81.73**	0.01	**0.01**	0.00	**+0.01**

Representation-level evidence. Figure 4(a) visualizes fusion-layer representations on SNLI-VE using UMAP. GU and layer interpolation exhibit similar broad layouts, consistent with the limited scope of their single-layer edits. The quantitative analysis in Appendix F.2 shows that both edits retain high similarity to the original TI-T representations, while GU achieves larger error-margin gains and corrects more errors with a small increase in regressions. With its TI-T F1, these results support targeted task-relevant correction with substantial preservation of the representation structure.

Overall, the main results show consistent performance gains under complete missing-image inference, with GU ranking first among the methods in Table 2 across all three datasets. Its advantage lies in correcting harmful learned dependencies within the model, rather than relying on data-side modality recovery or adaptive routing when insufficient cross-modal information remains. Comparisons with more baselines are in Appendix F.4.

6.3 Robustness–preservation trade-off across model families

A key question is whether improving missing-modality robustness sacrifices full-modality capability. Table 3 addresses this robustness–preservation trade-off by comparing TI-T and TI-TI performance across model families. GU improves TI-T performance across model families while keeping TI-TI performance nearly unchanged, indicating that the edited models do not collapse into text-only solutions. Compared with layer interpolation, GU usually yields larger TI-T gains and smaller TI-TI degradation, suggesting that subspace-level editing better preserves useful multimodal structure than direct parameter mixing. Complete per-variant results are provided in Appendix F.3.

The gain magnitude varies across architectures. GU is strongest for fusion models, where cross-modal dependencies are relatively localized in explicit fusion layers. Its gains are smaller for CLIP-style and BLIP architectures, where the relevant dependencies can be more distributed. Thus, GU is best viewed as a localized robustness correction method. It improves the robustness–preservation trade-off most when the harmful dependency is concentrated in the selected layer.

6.4 Cost analysis

Beyond performance, we also assess GU's computational cost and deployment efficiency. We compare GU, post-hoc LoRA, and Dual CE on SNLI-VE using Fusion Cross-Attention, CLIP Mean Pooling, and BLIP-2 under the same hardware/software environment and evaluation protocol. LoRA starts from the same TI checkpoint as GU and adapts the same target layer with rank 8. It trains on missing-image inputs for 4 epochs at learning rate 10^{-4}, selecting the best checkpoint on the development set. Dual CE retrains from scratch with joint full-input and missing-image supervision.

GU requires only a one-time edit before deployment, with no gradient updates to the TI checkpoint. As shown in Table 4, this stage takes 230–858 seconds, reducing wall-clock time by 57–88% relative to post-hoc LoRA. The resulting model introduces no additional parameters or forward-pass FLOPs relative to the original TI model. Inference uses only the edited checkpoint, without retaining the unimodal reference. For the complete offline comparison, costs start from an available TI checkpoint, excluding its original training cost. GU's total cost includes matched reference training and editing, whereas Dual CE includes its full retraining cost. The two GU phases run sequentially, so total runtime is their sum and peak memory is the larger of their two peaks. Even including reference training, GU uses 7–22% less wall-clock time and 16–62% less peak GPU memory than Dual CE across all three architectures.

Table 4: Offline cost and robustness–preservation trade-off on SNLI-VE. Including reference training, GU uses less time and peak GPU memory than Dual CE.

Architecture	Method / Phase	Wall-clock (s)	Peak Mem. (GiB)	ΔT	ΔTI
Fusion CA	Post-hoc LoRA	1812.30	2.090	+6.02	-1.62
	Dual CE	2592.03	13.954	+13.57	-0.94
	T Reference	1849.09	5.241	–	–
	GU Edit	297.19	1.744	–	–
	GU (Total)	2146.28	5.241	+8.06	-0.11
CLIP MP	Post-hoc LoRA	1910.39	4.423	+1.03	-3.66
	Dual CE	2580.61	3.684	+2.16	-1.40
	T Reference	1773.52	3.099	–	–
	GU Edit	230.17	3.104	–	–
	GU (Total)	2003.69	3.104	+1.43	-0.97
BLIP-2	Post-hoc LoRA	2000.73	16.930	+0.26	-0.68
	Dual CE	3079.53	32.857	+4.41	-0.53
	T Reference	1993.90	24.957	–	–
	GU Edit	857.83	16.126	–	–
	GU (Total)	2851.73	24.957	+0.01	+0.01

Table 4 also reports ΔT and ΔTI relative to the original TI model. GU combines lower total cost than Dual CE with better full-modality preservation. On Fusion and CLIP, it also achieves larger TI-T gains and smaller TI-TI losses than LoRA, although reference training makes its total offline runtime longer; on BLIP-2, the single-layer edit mainly preserves the original performance. Overall, GU provides an efficient post-training robustness correction with low offline cost, no additional inference overhead, and strong full-modality preservation.

6.5 Ablation studies

We ablate three key factors in GU: the geodesic coefficient η, the subspace rank k, and the target layer. As shown in Figure 4(b), GU is relatively stable across a range of η and k values, while the target layer plays a more decisive role. Varying η from 0.25 to 1.0 causes only moderate changes in missing-modality performance and full-modality preservation, while ranks from 2 to 32 yield similar TI-T performance with small TI-TI fluctuations. We therefore select hyperparameters based on the development-set trade-off between TI-T improvement and TI-TI preservation.

Editing most layers yields limited gains, whereas applying GU to the selected cross-modal interaction layer `attn_v_from_t_in` produces the largest TI-T improvement while maintaining strong TI-TI performance. This aligns with Section 4, where subspace rotation and representational harm concentrate in cross-modal interaction stages. This result further supports the localized nature of missing-modality degradation, showing that a single well-chosen Grassmannian subspace edit can recover robustness without broadly modifying the model.

7 Conclusion

We systematically characterize a deployment-time failure mode in which full-modality multimodal models can underperform unimodal models when a modality is absent at inference. We show that this degradation is strongly associated with the principal subspace rotations of the cross-modal interaction layers and task-aware representational harm. Motivated by these findings, we propose Geodesic Unlearning, a lightweight post-training method that realigns a selected principal subspace, viewed as a point on the Grassmannian, toward a unimodal reference, with provably optimal alignment under a fixed subspace-distance budget. Experiments across model families demonstrate a favorable robustness–preservation trade-off, with further support from multi-choice VQA and missing-text evaluation. Overall, our theoretical and empirical results support a geometric view of the missing-modality problem and suggest localized subspace editing as a practical correction strategy.

Acknowledgments and Disclosure of Funding

This research was partially supported by the U.S. National Science Foundation under Award No. 2310260. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors acknowledge the Center for Research Computing at Rice University for providing computational resources. The authors declare no competing interests.

References

[1] Hritam Basak and Zhaozheng Yin. Forget more to learn more: Domain-specific feature unlearning for semi-supervised and unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 130–148. Springer, 2024.

[2] Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

[4] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. doi: 10.1109/SP.2015.35.

[5] Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gjwhDHeAsz.

[6] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[7] Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations, 2021. URL https://arxiv.org/abs/2004.03744.

[8] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072, 2025.

[9] Siyi Du, Xinzhe Luo, Declan O'regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PWhDUWRVhM.

[10] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[11] Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in neural information processing systems*, 31, 2018.

[12] Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of VAEs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=s4MwstmB8o.

[13] Jack Geraghty, Andrew Hines, and Fatemeh Golpayegani. Learning to associate: Multimodal inference with fully missing modalities. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–48, 2025.

[14] Aditya Golatkar, Alessandro Achille, Luca Zancato, Yu-Xiang Wang, Ashwin Swaminathan, and Stefano Soatto. Cpr: Retrieval augmented generation for copyright protection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12374–12384, 2024.

[15] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

[16] Xiaofeng Han, Shunpeng Chen, Zenghuang Fu, Zhe Feng, Lue Fan, Dong An, Changwei Wang, Li Guo, Weiliang Meng, Xiaopeng Zhang, et al. Multimodal fusion and vision-language models: A survey for robot vision. *Information Fusion*, page 103652, 2025.

[17] Jun He, Laura Balzano, and John Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.

[18] Zelin He, Sarah Alnegheimish, and Matthew Reimherr. Harnessing vision-language models for time series anomaly detection. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i26.39319. URL https://doi.org/10.1609/aaai.v40i26.39319.

[19] 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, 2021. URL https://arxiv.org/abs/2106.09685.

[20] Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, 37:67446–67466, 2024.

[21] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.

[22] Dina Hussein, Taha Belkhouja, Ganapati Bhat, and Jana Doppa. Sensor-aware data imputation for time-series machine learning on low-power wearable devices. *ACM Transactions on Design Automation of Electronic Systems*, 30(1):1–27, 2024.

[23] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

[24] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

[25] Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. Approximate domain unlearning for vision-language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 21805–21833. Curran Associates, Inc., 2025. doi: 10.52202/085713-0736. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1f6af963e891e7efa229c24a1607fa7f-Paper-Conference.pdf.

[26] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[27] Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Liang, Matthew Blaschko, and Maarten De Vos. Balancing multimodal training through game-theoretic regularization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 167953–167982. Curran Associates, Inc., 2025. doi: 10.52202/085713-5600. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f5784de266e45bd09333721b0a602b61-Paper-Conference.pdf.

[28] Brett Koonce. Resnet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72. Springer, 2021.

[29] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[30] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22214–22224, October 2023.

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[32] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *International Journal of Computer Vision*, 134(3):1–31, 2026. doi: 10.1007/s11263-025-02667-1.

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[34] Shen-Huan Lyu, Lu Wang, and Zhi-Hua Zhou. Improving generalization of deep neural networks by leveraging margin distribution. *Neural Networks*, 151:48–60, 2022.

[35] 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, pages 2302–2310, 2021.

[36] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117 (40):24652–24663, 2020.

[37] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[38] Parjanya Prajakta Prashant, Kaustubh Ponkshe, and Babak Salimi. Tokenswap: A lightweight method to disrupt memorized sequences in llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[39] Sahar Rajabi, Nayeema Nonta, and Sirisha Rambhatla. Subtrack++ : Gradient subspace tracking for scalable llm training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 31584–31613. Curran Associates, Inc., 2025. doi: 10.52202/085713-1060. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2d62cb71e87ae340e3ab0e874befcbc2-Paper-Conference.pdf.

[40] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

[41] Thomas Robert, Mher Safaryan, Ionut-Vlad Modoranu, and Dan Alistarh. LDAdam: Adaptive optimization from low-dimensional gradient statistics. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Zkp1GuHerF.

[42] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[43] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pages 56–67. SPIE, 2005.

[44] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[45] Songyuan Sui, Hongyi Liu, Serena Liu, Li Li, Soo-Hyun Choi, Rui Chen, and Xia Hu. Chain-of-query: Unleashing the power of LLMs in SQL-aided table understanding via multi-agent collaboration. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 957–986, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.53. URL https://aclanthology.org/2025.ijcnlp-long.53/.

[46] Songyuan Sui, Zhanqi Deng, Yamin Zhuang, Serena Liu, Leisheng Yu, Alfredo Costilla Reyes, and Xia Hu. Eclad: An edge-cloud collaborative agentic framework for interpretable anomaly detection in predictive maintenance. In *Companion Proceedings of the ACM Web Conference 2026*, WWW Companion '26, pages 164–167, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400723087. doi: 10.1145/3774905.3793127. URL https://doi.org/10.1145/3774905.3793127.

[47] Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15173–15182, 2024.

[48] Ye Sun, Hao Zhang, Tiehua Zhang, Xingjun Ma, and Yu-Gang Jiang. Unseg: One universal unlearnable example generator is enough against all image segmentation. *Advances in Neural Information Processing Systems*, 37:79168–79193, 2024.

[49] Yuhang Sun, Zhizhong Liu, Quan Z Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110:102454, 2024.

[50] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023.

[51] Hao Wang, Zhengnan Li, Zhichao Chen, Xu Chen, Shuting He, Guangyi Liu, Haoxuan Li, and Zhouchen Lin. Iterative missing data imputation with model form adaptation and non-missing feature supervision. In *The thirty-ninth annual conference on neural information processing systems*, 2025.

[52] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15878–15887, 2023.

[53] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 216–226. Springer, 2023.

[54] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

[55] Zhenyi Wang, Yan Li, Li Shen, and Heng Huang. A unified and general framework for continual learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BE5aK0ETbp.

[56] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2024.

[57] Jiageng Wen, Shengjie Zhao, Bing Li, Jiafeng Huang, Kenan Ye, and Hao Deng. SiMO: Single-modality-operable multimodal collaborative perception. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=h0iRgjTmVs.

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

[59] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[60] Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. Distilled prompt learning for incomplete multimodal survival prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5102–5111, 2025.

[61] Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2575–2584. IEEE, 2023.

[62] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[63] Yanwu Yang, Hairui Chen, Zhikai Chang, Yang Xiang, Chenfei Ye, and Ting Ma. Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *Medical Imaging With Deep Learning*, pages 1006–1018. PMLR, 2024.

[64] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16416–16424, 2024.

[65] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37: 98782–98805, 2024.

[66] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.49. URL https://aclanthology.org/2023.emnlp-main.49/.

[67] Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, and Changsheng Xu. Unlearnable clusters: Towards label-agnostic unlearnable examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3984–3993, 2023.

[68] Yuan Zhang, Hu Wang, David Butler, Minh-Son To, Jodie Avery, M Louise Hull, and Gustavo Carneiro. Distilling missing modality knowledge from ultrasound for endometriosis diagnosis with magnetic resonance images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

[69] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, 44(1):4–18, 2024.

[70] Di Zhao, Jingfeng Zhang, Hongsheng Hu, Philippe Fournier-Viger, Gillian Dobbie, and Yun Sing Koh. Unlearning during training: Domain-specific gradient ascent for domain generalization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9ufS5Jl0O0.

[71] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM computing surveys*, 56(9):1–36, 2024.

[72] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: memory-efficient llm training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, pages 61121–61143, 2024.

[73] Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. MoRA: Missing modality low-rank adaptation for visual recognition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZgQnIPG4uV.

[74] Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities. In *European Conference on Computer Vision*, pages 146–165. Springer, 2024.

Appendix Contents

A Limitations **18**

B Broader impacts **18**

C Supplementary materials for empirical analysis **18**

C.1 More results on missing-modality degradation 18

C.2 More results on parameter subspace rotation 18

C.3 More results on representational harm under missing modalities 18

D Methodology details **21**

D.1 Grassmannian geodesic construction 21

D.2 Geometric rationale and path comparison 21

D.3 Implementation notes 22

E Supplementary materials for experimental setup **22**

E.1 Dataset 22

E.2 Baselines 22

E.3 Compute resources 24

E.4 Implementation 24

F Supplementary experimental results **26**

F.1 Performance stability analysis 26

F.2 More details on representation-level evidence 26

F.3 More results on robustness–preservation trade-off 27

F.4 Comparisons with more baselines 27

F.5 Controlled interventions for geometric correction 30

F.6 Generalization to multiple-choice VQA 31

F.7 Complementary evaluation under missing-text inputs 31

A Limitations

Despite the importance of our work in highlighting and addressing the structural causes of missing-modality degradation in multimodal learning, there are limitations worth noting. First, our evaluation is limited to classification tasks, including multiple-choice visual question answering. Extending GU to free-form multimodal generation is an important direction for future work. Second, concerns have been raised about rendered text in Hateful Memes and hypothesis-side artifacts in SNLI-VE. We address these concerns on the affected benchmarks through targeted preprocessing and complementary controls. Further evaluations on more datasets without these specific artifacts reproduce both the degradation and GU's improvements, confirming that neither finding requires their presence. Finally, although we cover fusion models, two-tower models, and VLMs across diverse domains, due to resource constraints, we do not conduct large-scale experiments on substantially larger models.

B Broader impacts

This work improves the robustness of multimodal models to missing modality inputs, which is important in real-world settings where inputs may be incomplete due to sensor failure, privacy constraints, or deployment limitations. Such improvements may benefit applications like medical decision support and multimodal content understanding. However, increased robustness may encourage deployment in settings where incorrect predictions can still occur. In addition, robustness does not eliminate biases present in training data.

C Supplementary materials for empirical analysis

C.1 More results on missing-modality degradation

The main text summarizes missing-modality degradation using aggregated bar plots. Table 5 provides complete F1 scores across model variants and datasets under full-modality evaluation (TI-TI), unimodal text evaluation (T-T), and deployment-time missing-image evaluation (TI-T). Results are mean $\pm$ sample standard deviation over three independent runs. IX and SV denote IU-XRay and SNLI-VE, respectively.

The reported mean scores support the trend in Section 4.2: TI-TI $>$ T-T $>$ TI-T holds in 26 of 27 settings, and T-T $>$ TI-T holds in all 27. The only non-strict case is strong-encoder Fusion Cross-Attention on HM, where TI-TI $=$ T-T $=$ 66.93 $>$ TI-T $=$ 66.85. The effect is especially pronounced in models with explicit cross-modal interaction, such as cross-attention fusion on SNLI-VE, where the missing-modality gap is among the largest. These results support the view that deployment-time missing-modality degradation is associated not only with absent input information but also with cross-modal dependencies formed during multimodal training.

C.2 More results on parameter subspace rotation

To complement the main-text analysis of interaction-layer geometry, we provide a layer-wise comparison of fusion models on SNLI-VE. Table 7 reports input/output subspace overlap, mean principal angle, and Frobenius distance between each TI layer and its matched T reference.

Subspace shifts are nonuniform across layers. Projection and cross-modal interaction layers generally exhibit larger principal-angle deviations and lower subspace overlap than classifier layers. Cross-attention modules show large input- and output-side angle shifts, consistent with the main-text finding that interaction-layer geometry most strongly tracks the missing-modality gap. By contrast, the final classifier has near-perfect overlap and a near-zero output-side angle, suggesting that the dominant parameter-space changes arise before the final prediction layer.

These results support interaction-layer subspace rotation as a structural diagnostic for missing-modality degradation and a criterion for selecting localized GU editing targets.

C.3 More results on representational harm under missing modalities

The main text summarizes task-aware representational harm at the stage level. Table 6 provides layer-wise results for fusion models on SNLI-VE under missing-image evaluation. It reports the

Table 5: F1 performance under three training–testing regimes. TI-T consistently underperforms T-T across model families and datasets.

Family	Dataset	Variant	TI-TI	T-T	TI-T	Δ_{miss}
Fusion	HM	Concat	69.46 ± 0.32	68.99 ± 0.25	68.96 ± 0.44	0.50 ± 0.14
	HM	Gate	67.98 ± 0.28	67.86 ± 0.23	67.57 ± 0.38	0.41 ± 0.12
	HM	Cross-Attn	68.82 ± 0.48	68.41 ± 0.40	57.63 ± 0.88	11.19 ± 0.45
	IX	Concat	86.41 ± 0.22	81.94 ± 0.18	80.70 ± 0.38	5.71 ± 0.18
	IX	Gate	82.79 ± 0.20	81.18 ± 0.17	80.76 ± 0.33	2.03 ± 0.15
	IX	Cross-Attn	84.57 ± 0.28	80.97 ± 0.23	80.00 ± 0.47	4.57 ± 0.22
	SV	Concat	86.22 ± 0.07	81.84 ± 0.05	73.04 ± 0.24	13.18 ± 0.18
	SV	Gate	86.53 ± 0.08	82.31 ± 0.06	71.70 ± 0.29	14.83 ± 0.22
	SV	Cross-Attn	86.64 ± 0.10	81.85 ± 0.08	66.35 ± 0.38	20.29 ± 0.30
Fusion (strong)	HM	Concat	67.49 ± 0.30	67.12 ± 0.24	66.50 ± 0.46	0.99 ± 0.18
	HM	Gate	67.12 ± 0.34	66.85 ± 0.27	65.51 ± 0.53	1.61 ± 0.22
	HM	Cross-Attn	66.93 ± 0.22	66.93 ± 0.18	66.85 ± 0.24	0.08 ± 0.03
	IX	Concat	81.51 ± 0.21	80.35 ± 0.16	78.06 ± 0.36	3.45 ± 0.17
	IX	Gate	79.93 ± 0.18	79.79 ± 0.14	79.24 ± 0.25	0.69 ± 0.08
	IX	Cross-Attn	80.14 ± 0.24	79.46 ± 0.19	77.78 ± 0.38	2.36 ± 0.16
	SV	Concat	76.44 ± 0.06	75.99 ± 0.05	72.78 ± 0.19	3.66 ± 0.14
	SV	Gate	72.72 ± 0.07	71.51 ± 0.06	67.65 ± 0.24	5.07 ± 0.18
	SV	Cross-Attn	79.71 ± 0.08	79.23 ± 0.07	73.69 ± 0.27	6.02 ± 0.20
CLIP	HM	Late Concat	67.31 ± 0.27	67.21 ± 0.22	66.58 ± 0.41	0.73 ± 0.16
	HM	Mean Pooling	67.14 ± 0.20	66.95 ± 0.17	66.87 ± 0.24	0.27 ± 0.05
	IX	Late Concat	80.85 ± 0.19	80.40 ± 0.15	79.45 ± 0.29	1.40 ± 0.12
	IX	Mean Pooling	78.00 ± 0.20	77.97 ± 0.16	74.30 ± 0.38	3.70 ± 0.20
	SV	Late Concat	78.05 ± 0.06	78.01 ± 0.05	72.77 ± 0.21	5.28 ± 0.16
	SV	Mean Pooling	76.62 ± 0.06	76.52 ± 0.05	74.15 ± 0.18	2.47 ± 0.13
VLM (BLIP-2)	HM	BLIP-2	68.85 ± 0.38	66.58 ± 0.31	66.21 ± 0.59	2.64 ± 0.24
	IX	BLIP-2	81.74 ± 0.26	79.38 ± 0.22	78.00 ± 0.46	3.74 ± 0.23
	SV	BLIP-2	81.72 ± 0.18	77.51 ± 0.07	70.64 ± 0.70	11.08 ± 0.66

error–correct harm gap Δ_{harm} and violation-rate gap Δ_{viol}. Larger gaps indicate that error samples occupy more task-inconsistent regions than correct samples.

Representational harm is nonuniform across layers. Harm signals become pronounced around multimodal interaction stages and often persist into the classifier input. This complements the parameter-subspace analysis in Table 7. Image projection layers can exhibit large subspace rotations, but their harm scores are zero under missing-image evaluation because the image modality is unavailable. By contrast, cross-modal interaction layers show both substantial subspace rotation and nontrivial representational harm. These results motivate selecting target layers using both parameter subspace rotation and task-aware harm, rather than principal angles alone.

Together, these findings support the main-text interpretation that interaction-layer geometry is accompanied by task-relevant representation shifts under missing-modality inputs.

Table 6: Layer-wise task-aware representational harm for fusion models on SNLI-VE. Harm gaps appear at cross-modal interaction stages and often persist near the classifier.

Variant	Layer	Δ_{harm}	Δ_{viol}
Concat	text_proj_out	0.5187	0.6170
	image_proj_out	0.0000	0.0000
	fusion_mlp0_concat_input	0.5187	0.6170
	fusion_mlp3_output	0.5617	0.7308
	head_last_linear_input	0.5617	0.7308
Gate	text_proj_out	0.3881	0.5735
	image_proj_out	0.0000	0.0000
	fusion_mlp0_concat_input	0.3881	0.5735
	fusion_mlp3_h_base	0.2373	0.5992
	gate_mlp0_gate	0.1066	0.5617
	fusion_pre_ln	0.2996	0.6127
	fusion_output	0.2996	0.6113
	head_last_linear_input	0.2996	0.6113
Cross-Attn	text_proj_out	0.4585	0.4339
	image_proj_out	0.0000	0.0000
	attn_t_from_v_in_proj_to_text_pooled	0.4134	0.4464
	attn_t_from_v_out_proj_to_text_pooled	0.4134	0.4464
	attn_v_from_t_in_proj_to_image_pooled	0.3673	0.4557
	attn_v_from_t_out_proj_to_image_pooled	0.3673	0.4557
	ffn_t_0_to_text_pooled	0.4134	0.4464
	ffn_t_3_to_text_pooled	0.4134	0.4464
	ffn_v_0_to_image_pooled	0.3673	0.4557
	ffn_v_3_to_image_pooled	0.3673	0.4557
	combine_mlp0_concat_input	0.3896	0.4420
	combine_mlp3_output	0.4445	0.5868
	head_last_linear_input	0.4445	0.5868

Table 7: Layer-wise parameter subspace rotation for fusion models on SNLI-VE. Projection and cross-modal interaction layers generally exhibit larger rotations than the final classifier.

Variant	Layer	Overlap (In / Out)	Mean Angle (In / Out)	Fro. Dist.
Concat	text_proj	0.9666 / 0.9666	8.3333 / 8.3150	1.4884
	image_proj	0.8368 / 0.8409	19.2640 / 18.7282	2.5745
	fusion_mlp0	0.9307 / 0.9314	11.4472 / 11.3110	1.8050
	fusion_mlp3	0.9578 / 0.9573	4.2398 / 4.2895	1.3692
	head_last_linear	0.9991 / 1.0000	1.7072 / 0.0000	0.0511
Gate	text_proj	0.9715 / 0.9714	7.7292 / 7.7325	1.4133
	image_proj	0.8615 / 0.8656	17.8397 / 17.2876	2.4042
	fusion_mlp0	0.9129 / 0.9126	12.3658 / 12.2981	1.8239
	fusion_mlp3	0.9249 / 0.9254	8.3895 / 8.2923	1.4560
	gate_mlp0	0.9408 / 0.9425	10.8734 / 10.5926	2.0828
	head_last_linear	0.9994 / 1.0000	1.3227 / 0.0000	0.0367
Cross-Attn	text_proj	0.9656 / 0.9658	8.1285 / 8.0749	1.4191
	image_proj	0.8777 / 0.8818	16.1052 / 15.5300	2.3401
	attn_t_from_v_in_proj	0.9152 / 0.9127	13.3996 / 13.8179	2.3295
	attn_t_from_v_out_proj	0.8701 / 0.8666	17.2543 / 17.6767	1.4386
	attn_v_from_t_in_proj	0.8845 / 0.8796	16.0031 / 16.5666	2.4871
	attn_v_from_t_out_proj	0.9315 / 0.9312	10.2142 / 10.2819	0.9027
	ffn_t_0	0.9873 / 0.9870	5.2286 / 5.3021	1.1664
	ffn_t_3	0.9862 / 0.9860	5.0457 / 5.0703	1.0904
	ffn_v_0	0.9771 / 0.9767	6.4555 / 6.4909	1.2853
	ffn_v_3	0.9881 / 0.9882	4.6682 / 4.6642	1.0396
	combine_mlp0	0.9638 / 0.9642	8.4807 / 8.4070	1.5600
	combine_mlp3	0.9471 / 0.9464	5.3530 / 5.4254	1.0109
	head_last_linear	0.9988 / 1.0000	1.9483 / 0.0000	0.0501

D Methodology details

D.1 Grassmannian geodesic construction

Intuitively, the construction pairs the principal directions of the multimodal and unimodal subspaces and rotates each multimodal direction toward its paired reference direction. A single edit strength specifies the same fraction of angular correction for every pair. The resulting basis is then expressed in the original TI coordinates for reconstruction with the retained output factor, principal coefficients, and residual. This section provides the explicit Grassmannian geodesic used in Eq. (3). Let $V_{\mathrm{TI},k}, V_{\mathrm{T},k} \in \mathbb{R}^{d_i \times k}$ be orthonormal bases of the top-k input subspaces of the multimodal model and the matched unimodal reference. We first compute the compact SVD of their subspace overlap:

$$V_{\mathrm{TI},k}^{\top} V_{\mathrm{T},k} = P \cos(\Theta) Q^{\top}, \tag{6}$$

where the diagonal entries $\theta_i \in [0, \pi/2]$ of Θ are the principal angles between the two subspaces. We then form the aligned principal bases

$$A = V_{\mathrm{TI},k} P, \qquad B = V_{\mathrm{T},k} Q. \tag{7}$$

The geodesic direction is computed as

$$Y = (B - A\cos\Theta)(\sin\Theta)^{\dagger}, \tag{8}$$

where $(\cdot)^{\dagger}$ denotes the Moore–Penrose pseudoinverse. For $\theta_i = 0$, the common direction is unchanged and the corresponding column of Y is zero. The geodesic basis in principal coordinates and its expression in the original TI singular-vector coordinates are

$$\widetilde{V}(\eta) = A\cos(\eta\Theta) + Y\sin(\eta\Theta), \qquad V(\eta) = \widetilde{V}(\eta) P^{\top}, \quad \eta \in [0, 1]. \tag{9}$$

Here $\widetilde{V}(0) = A$ and $\widetilde{V}(1) = B$. Multiplication by $P^{\top}$ preserves the subspace and orthonormality while expressing the basis in the TI coordinates used for reconstruction. Consequently, $V(0) = V_{\mathrm{TI},k}$ and $V(1) = V_{\mathrm{T},k} Q P^{\top}$, so the path starts at the original TI basis and ends in the reference subspace.

D.2 Geometric rationale and path comparison

The edited object. The top-k right singular subspace specifies the dominant input directions read by the layer. Sections 4.3 and 4.4 motivate targeting this orientation through parameter deviation and task-aware representation harm. Its Grassmannian representation identifies all orthonormal bases spanning the same subspace. GU retains the original output factor $U_{\mathrm{TI},k}$, principal coefficients $\Sigma_{\mathrm{TI},k}$, and residual R_{TI}; direct weight interpolation changes these components together.

Edit-budget optimality. The geometric question is how close an edited subspace can get to the reference while limiting its departure from the original multimodal subspace. The edit budget specifies this allowed departure. A shortest path uses the budget fully toward alignment: every unit of movement reduces the remaining distance to the reference by one unit.

Let $S_{\mathrm{TI}} = \mathrm{span}(V_{\mathrm{TI},k})$ and $S_{\mathrm{T}} = \mathrm{span}(V_{\mathrm{T},k})$. We use the canonical Grassmann geodesic distance [10],

$$d(S_1, S_2) = \left(\sum_{i=1}^{k} \theta_i(S_1, S_2)^2 \right)^{1/2}, \qquad D = d(S_{\mathrm{TI}}, S_{\mathrm{T}}). \tag{10}$$

This distance defines the edit budget; the mean principal angle in Section 4.3 is the diagnostic summary. The subspace path in Eq. (9) can be written intrinsically as

$$S_\eta = \mathrm{Exp}_{S_{\mathrm{TI}}}\left(\eta\, \mathrm{Log}_{S_{\mathrm{TI}}}(S_{\mathrm{T}})\right), \qquad 0 \le \eta \le 1, \tag{11}$$

where Log denotes a minimizing logarithm; when it is nonunique, a minimizing branch is fixed. Each principal direction advances through $\eta\theta_i$, leaving $(1-\eta)\theta_i$ to the reference. Hence $d(S_{\mathrm{TI}}, S_\eta) = \eta D$ and $d(S_\eta, S_{\mathrm{T}}) = (1-\eta)D$. For any candidate S satisfying $d(S_{\mathrm{TI}}, S) \le \eta D$, the triangle inequality gives

$$d(S, S_{\mathrm{T}}) \ge D - d(S_{\mathrm{TI}}, S) \ge (1-\eta)D. \tag{12}$$

The geodesic point S_η attains this bound, proving Eq. (5). Thus, for a fixed subspace-distance budget, GU makes the greatest possible progress toward the matched reference. This optimality concerns the geometric path; it does not assert optimal task performance.

Comparison with aligned linear interpolation. Using the same principal bases A, B, consider the extrinsic path

$$L(\eta) = \mathrm{polar}((1-\eta)A + \eta B)\,, \tag{13}$$

where $\mathrm{polar}(X) = X(X^\top X)^{-1/2}$ is the orthonormal polar factor. Write a_i, b_i, y_i for corresponding columns of A, B, Y. Since $b_i = \cos\theta_i\, a_i + \sin\theta_i\, y_i$, the ith interpolated column is

$$\ell_i(\eta) = \frac{(1-\eta+\eta\cos\theta_i)a_i + \eta\sin\theta_i\, y_i}{\sqrt{(1-\eta+\eta\cos\theta_i)^2 + \eta^2\sin^2\theta_i}}. \tag{14}$$

The unnormalized columns are mutually orthogonal, so the polar step normalizes each separately. Its displacement from a_i within the corresponding principal plane is therefore

$$\phi_i(\eta) = \mathrm{atan2}(\eta\sin\theta_i,\ 1-\eta+\eta\cos\theta_i)\,. \tag{15}$$

In general, $\phi_i(\eta) \neq \eta\theta_i$. The distinction is the meaning of the interpolation coefficient: mixing all aligned basis vectors with the same coefficient does not generally advance them through the same fraction of their respective angles. GU gives η a consistent geometric meaning across directions, advancing every principal angle proportionally at constant geodesic speed. For weight reconstruction, the corresponding linear-path basis is $L(\eta)P^\top$, using the same TI coordinates as GU.

D.3 Implementation notes

In implementation, the singular values of $V_{\mathrm{TI},k}^\top V_{\mathrm{T},k}$ are clamped to $[-1, 1]$ before applying $\arccos$ to compute the principal angles. For angles close to zero, the corresponding entries of $(\sin\Theta)^\dagger$ are stabilized with a small numerical constant. After computing $V(\eta)$, the edited layer is reconstructed using Eq. (4) in the main text:

$$W' = R_{\mathrm{TI}} + U_{\mathrm{TI},k}\Sigma_{\mathrm{TI},k}V(\eta)^\top. \tag{16}$$

The conversion in Eq. (9) ensures that $W' = W_{\mathrm{TI}}$ when $\eta = 0$. All other layers remain unchanged.

E Supplementary materials for experimental setup

E.1 Dataset

Table 8: Dataset split statistics used in our experiments.

Split	Hateful Memes	SNLI-VE	IU-XRay
Train	8,500	353,482	2,928
Dev	500	11,898	366
Test	1,000	11,937	366

Our main experiments use three text-image multimodal benchmarks: *Hateful Memes* [26], *SNLI-VE* [59], and *IU-XRay* [6]. *Hateful Memes* is a social multimodal classification benchmark that requires jointly reasoning over meme images and associated text. *SNLI-VE* is a visual entailment benchmark that pairs images with textual hypotheses. *IU-XRay* is a medical vision-language dataset constructed from chest X-ray images and radiology reports. For consistency, they are formulated as binary classification tasks. For *SNLI-VE*, which originally comprises entailment, contradiction, and neutral classes, we remove neutral examples. This maintains a consistent binary classification protocol across datasets while mitigating the impact of documented labeling issues associated with the neutral class [15, 7]. Table 8 summarizes the dataset splits used in our experiments. We use the development set for hyperparameter selection and report final results on the held-out test set.

E.2 Baselines

For experimental organization, we group the compared methods into multimodal-specific methods, general training and adaptation baselines, and parameter-editing controls. This grouping reflects each method's role and applicability in the evaluation, rather than a mutually exclusive taxonomy of prior work. Coverage follows each method's architectural requirements: SMIL, Flex-MoE, and DyMo are compared under a matched fusion backbone, while additional architecture-compatible methods extend the evaluation to other model families.

E.2.1 Multimodal-specific methods

SMIL, Flex-MoE, and DyMo. We evaluate SMIL [35], Flex-MoE [65], and DyMo [9] on all three main datasets using the unified RoBERTa–ResNet50 cross-attention fusion backbone. This choice follows the structural assumptions of their released implementations. SMIL couples Bayesian meta-learning with missing-modality feature reconstruction, requiring choices of reconstruction targets and auxiliary-network insertion points. Flex-MoE uses a missing-modality bank, experts, and specialized routers; DyMo uses recovered-modality candidates and inference-time modality selection. Extending these methods to CLIP-style or BLIP-2 models is possible, but would require method-specific changes to reconstruction targets, native fusion or Q-Former components, or routing and recovery modules. Such redesigns would alter the architecture, parameter budget, and optimization pipeline, adding design choices to matched-backbone comparisons. This setting provides a common backbone and training protocol while retaining each method's auxiliary components.

For SMIL, Flex-MoE, and DyMo, we use the official codebases released by the original papers and keep unspecified settings consistent with their default configurations. To ensure a fair comparison, we use the same number of training epochs across our models: 10 for Hateful Memes, 10 for IU-XRay, and 4 for SNLI-VE. We also match the main learning rates of these baselines to our unified cross-attention fusion setting. The only exception is the SMIL inner-loop learning rate, which we set to 1.0×10^{-3} to follow the official implementation.

MoRA and OGM-GE. MoRA [73] provides a missing-modality adaptation baseline on CLIP Late Concat and Mean Pooling across the three main datasets. Its modality-specific and shared low-rank updates couple paired text and vision encoder weights; applying this formulation to Fusion Concat would require new insertion points and a fusion-specific sharing rule. OGM-GE [37] provides a training-time modality-balancing baseline on Fusion Concat and both CLIP variants across the same datasets. CLIP Mean Pooling permits an exact decomposition of the full logit into text and image contributions. For nonlinear Late Concat, modality confidence is estimated through native unimodal forward paths while retaining the published discrepancy, gradient-modulation, and generalization-enhancement rules. Results are reported in Appendix F.4.1.

SiMO [57] uses a specialized collaborative-perception architecture and staged training. Porting it to our setting would replace the evaluated backbone and training pipeline, rather than provide a matched-backbone comparison.

E.2.2 General training and adaptation baselines

Dual CE. On SNLI-VE, we retrain Fusion Cross-Attention, CLIP Mean Pooling, and BLIP-2 from scratch with a shared classifier jointly supervising full-input and missing-image paths:

$$\mathcal{L}_{\text{Dual CE}} = \mathcal{L}_{\text{TI}} + 0.5\,\mathcal{L}_{\text{T}}.$$

Both terms use binary cross-entropy with logits, evaluated with full text–image inputs and with the image replaced by a null input, respectively. Dual CE+GU applies GU to the resulting checkpoint to test whether the localized edit complements joint supervision. The comparison is reported in Appendix F.4.2.

LoRA. For the SNLI-VE comparison on the same three architectures, post-hoc LoRA [19] starts from the same TI checkpoint as GU and adapts the same target layer on missing-image inputs. We use rank 8, train for 4 epochs at learning rate 10^{-4}, and select the best checkpoint on the development set. Performance and computational costs are reported in Section 6.4.

E.2.3 Parameter-editing controls

Weight (layer interpolation). For layer interpolation, we use the same target layer as Geodesic Unlearning for each architecture variant, as listed in Table 14. Given the selected multimodal parameter matrix W_{TI} and the corresponding unimodal reference parameter matrix W_{T}, layer interpolation constructs the edited parameter as

$$W' = (1-\alpha)W_{\text{TI}} + \alpha W_{\text{T}}.$$

We fix $\alpha = 0.5$ for all datasets and models. This baseline directly mixes parameter values, whereas GU edits only the selected Grassmannian subspace while preserving the residual parameter structure.

Linear (subspace interpolation). In the Fusion Cross-Attention controls on the three main datasets, Linear interpolates the aligned principal bases and applies polar retraction, as defined in Eq. (13). It uses the same target layer, matched checkpoints and reference, rank $k = 8$, and coefficient $\eta = 0.75$ as GU, while retaining the original TI output factor, principal coefficients, and residual through the reconstruction coordinates in Appendix D.2. These controls isolate the effect of the edited object and the subspace path; their results are reported in Appendix F.5.

E.3 Compute resources

All experiments can be run on a single NVIDIA A100 GPU with 80GB of memory. This includes training the multimodal and unimodal reference models, running Geodesic Unlearning and layer interpolation, and evaluating the missing-modality baselines. Geodesic Unlearning itself is applied as a post-training single-layer parameter edit and therefore adds negligible computational overhead compared with model training.

E.4 Implementation

E.4.1 Backbone configurations

For model families in Table 1, we use the following backbone configurations. Basic fusion models use RoBERTa and ResNet50, while the stronger fusion setting uses Qwen2.5-3B and OpenCLIP ViT-bigG-14. CLIP-style two-tower models use OpenCLIP ViT-L-14, and VLM experiments use BLIP-2 with a classifier head. For each TI model, we use a unimodal reference with the same architecture and pretrained parameterization. GU requires dimensionally and semantically corresponding layers; references without this correspondence would require an additional alignment map.

E.4.2 Hyperparameters

For the three binary benchmarks, standard TI training uses a shared classifier and binary cross-entropy with logits on full text–image inputs. Dual CE adds missing-image supervision as specified in Appendix E.2.

For all models used with Geodesic Unlearning, we train the underlying multimodal models with weight decay 1.0×10^{-2} and select the best checkpoint on the development set. The maximum number of training epochs is 10 for Hateful Memes, 10 for IU-XRay, and 4 for SNLI-VE. Tables 10–13 summarize the TI-mode learning rates used for each model family. The symbol "–" indicates that the module is not present in the corresponding configuration. For basic fusion models, no module-specific learning rates are used, and a single global learning rate of 3.0×10^{-5} is applied to all trainable parameters, including unfrozen encoder backbones.

E.4.3 Intervention details

For Geodesic Unlearning, we select the geodesic coefficient $\eta = 0.75$ and subspace rank $k = 8$ on the development set and keep both fixed across all datasets and model families. For each architecture variant, candidate layers are chosen from the interaction and nearby projection/classifier layers analyzed in Section 4. The target layer is selected on the development set using the angle–harm score in Eq. (1), which combines the principal-angle deviation between the TI model and the T reference with the task-aware harm gap under missing-modality evaluation. When a task-aware harm score is measured at the functional-stage level rather than at an individual layer, we assign the same stage-level harm score to all candidate layers within that stage. The test set is never used for target-layer selection. Table 14 lists the selected target layer for each architecture variant.

Table 9: Learning rates for missing-modality baselines. Unspecified hyperparameters follow the official codebase defaults.

Dataset	**GU**	SMIL Outer	SMIL Inner	Flex-MoE	DyMo
Hateful Memes	3.0×10^{-5}	3.0×10^{-5}	1.0×10^{-3}	3.0×10^{-5}	3.0×10^{-5}
SNLI-VE	3.0×10^{-5}	3.0×10^{-5}	1.0×10^{-3}	3.0×10^{-5}	3.0×10^{-5}
IU-XRay	3.0×10^{-5}	3.0×10^{-5}	1.0×10^{-3}	3.0×10^{-5}	3.0×10^{-5}

Table 10: Learning rates for basic fusion models. A single global learning rate is used for all trainable parameters.

Dataset	Concat	Cross-Attention	Gate
Hateful Memes	3.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
SNLI-VE	3.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
IU-XRay	3.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}

Table 11: Module-wise learning rates for fusion models with stronger encoders. Text/image denotes both `text_encoder.proj` and `image_encoder.proj`; bottleneck/head denotes both modules.

Dataset	Variant	Text/Image	Fusion	Bottleneck/Head
Hateful Memes	Concat	2.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
	Cross-Attention	2.0×10^{-5}	2.0×10^{-5}	3.0×10^{-5}
	Gate	1.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}
SNLI-VE	Concat	2.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
	Cross-Attention	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
	Gate	2.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
IU-XRay	Concat	2.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
	Cross-Attention	2.0×10^{-5}	2.0×10^{-5}	3.0×10^{-5}
	Gate	1.0×10^{-5}	2.0×10^{-5}	2.0×10^{-5}

Table 12: Module-wise learning rates for CLIP-style two-tower models. Text/image denotes both projection modules; bottleneck/head denotes both modules.

Dataset	Variant	Text/Image	Fusion	Bottleneck/Head
Hateful Memes	Late Concat	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
	Mean Pooling	1.0×10^{-5}	–	1.0×10^{-5}
SNLI-VE	Late Concat	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
	Mean Pooling	1.0×10^{-5}	–	1.0×10^{-5}
IU-XRay	Late Concat	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
	Mean Pooling	1.0×10^{-5}	–	1.0×10^{-5}

Table 13: Module-wise learning rates for BLIP-2 based VLM models. Q-Former denotes both `vlm_encoder.qformer_embeddings` and `vlm_encoder.backbone.qformer`; projection denotes both `vlm_encoder.backbone.language_projection` and `vlm_encoder.proj`.

Dataset	Q-Former	Projection	Bottleneck	Head
Hateful Memes	5.0×10^{-6}	5.0×10^{-6}	2.0×10^{-5}	2.0×10^{-5}
SNLI-VE	1.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}
IU-XRay	1.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}	3.0×10^{-5}

Table 14: Target layers edited by Geodesic Unlearning. We use $\eta = 0.75$ and $k = 8$ for all configurations.

Model Family	Variant	Target Layer
Fusion / Fusion (strong)	Concat	`fusion_mlp0`
Fusion / Fusion (strong)	Gate	`gate_mlp0`
Fusion / Fusion (strong)	Cross-Attention	`attn_v_from_t_in_proj`
Two-tower	Late Concat	`fusion_mlp0`
Two-tower	Mean Pooling	`text_proj`
VLM	BLIP-2 Classifier Head	`qformer_crossattn0_out`

Table 15: Supplementary results for Table 2. We report mean ± sample standard deviation over three independent runs. Underline and **bold** denote the second-best and best mean results, respectively.

Method	HM	IU-XRay	SNLI-VE
Null-token	57.63 ± 0.88	80.00 ± 0.47	66.35 ± 0.38
SMIL	56.60 ± 0.74	79.21 ± 0.68	$\underline{72.20} \pm 0.52$
Flex-MoE	47.58 ± 1.38	71.02 ± 0.82	70.74 ± 0.64
DyMo	48.82 ± 0.93	79.19 ± 0.77	61.30 ± 0.70
Interpolation	$\underline{58.50} \pm 0.91$	$\underline{80.14} \pm 0.49$	67.29 ± 0.61
GU (Ours)	$\mathbf{60.88} \pm 0.73$	$\mathbf{81.41} \pm 0.63$	$\mathbf{74.41} \pm 0.77$

Table 16: Supplementary robustness–preservation results for Table 3 on SNLI-VE. Results are mean ± sample standard deviation over three independent runs.

Metric	Method	Fusion	Fusion (str.)	CLIP	VLM
TI-T	Interp.	71.34 ± 0.46	72.34 ± 0.38	74.50 ± 0.44	70.65 ± 0.79
	GU	$\mathbf{73.86} \pm 0.56$	$\mathbf{73.06} \pm 0.43$	$\mathbf{74.68} \pm 0.52$	$\mathbf{70.65} \pm 0.68$
TI-TI	Interp.	86.33 ± 0.18	75.16 ± 0.37	76.13 ± 0.38	81.72 ± 0.30
	GU	$\mathbf{86.37} \pm 0.14$	$\mathbf{75.95} \pm 0.31$	$\mathbf{76.70} \pm 0.31$	$\mathbf{81.73} \pm 0.18$
ΔT	Interp.	0.98 ± 0.30	0.96 ± 0.22	1.04 ± 0.35	0.01 ± 0.10
	GU	$\mathbf{3.49} \pm 0.26$	$\mathbf{1.69} \pm 0.20$	$\mathbf{1.22} \pm 0.33$	$\mathbf{0.01} \pm 0.04$
ΔTI	Interp.	-0.13 ± 0.11	-1.13 ± 0.31	-1.21 ± 0.34	0.00 ± 0.13
	GU	$-\mathbf{0.09} \pm 0.06$	$-\mathbf{0.34} \pm 0.24$	$-\mathbf{0.64} \pm 0.25$	$+\mathbf{0.01} \pm 0.04$

F Supplementary experimental results

F.1 Performance stability analysis

F.1.1 Three-run means and standard deviations

Tables 15 and 16 supplement Tables 2 and 3, respectively, with mean ± sample standard deviation over three independent runs. For the family-level results, we first average the variants within each seed and then compute the mean and sample standard deviation across the three seed-level averages. The resulting standard deviations therefore describe variability across runs at the same aggregation level as the main-text means.

The three-run results retain GU's highest mean TI-T F1 on all three datasets in Table 15. The family-level results in Table 16 further show that GU matches or improves upon interpolation's mean missing-modality recovery and full-modality preservation across the evaluated families.

F.1.2 Ten-seed paired significance tests

We further validate the gains in Table 2 over ten matched seeds. For each dataset, GU is compared with the null-token baseline and the strongest competing method in that table: interpolation on HM and IU-XRay, and SMIL on SNLI-VE. We report the paired GU-minus-baseline F1 difference (ΔF1) as mean ± sample standard deviation. Statistical significance is assessed using two-sided exact paired sign-flip tests, with Holm correction across all six comparisons.

As shown in Table 17, all six comparisons remain significant after Holm correction ($p < 0.05$). Against the strongest competing method on each dataset, GU wins all 30 paired runs, with mean gains of $+2.41$, $+1.34$, and $+2.21$ F1 on HM, IU-XRay, and SNLI-VE, respectively. These results provide additional statistical support for the three-run main results reported separately above.

F.2 More details on representation-level evidence

We quantify the representation similarity and functional changes behind Figure 4(a) using the SNLI-VE Fusion Cross-Attention model. Table 18 reports mean ± sample standard deviation over

Table 17: Ten-seed paired tests for GU against Null and the strongest competing method in Table 2. All six comparisons remain significant after Holm correction.

Dataset	Comparison	10-seed paired ΔF1	GU wins	Raw p	Holm p
HM	GU−Null	$+3.19 \pm 1.14$	9/10	0.003906	0.011719
HM	GU−Interp.	$+2.41 \pm 0.77$	10/10	0.001953	0.011719
IX	GU−Null	$+1.51 \pm 1.71$	9/10	0.033203	0.033203
IX	GU−Interp.	$+1.34 \pm 0.45$	10/10	0.001953	0.011719
SV	GU−Null	$+8.09 \pm 0.22$	10/10	0.001953	0.011719
SV	GU−SMIL	$+2.21 \pm 0.68$	10/10	0.001953	0.011719

Table 18: Quantitative analysis of Figure 4(a) on SNLI-VE Fusion Cross-Attention. Both edits preserve high representation similarity, while GU yields larger task-level improvements.

Method	TI-T	CKA to orig.	Error-margin gain	Corr./regr. rate (%)
Orig. TI-T	66.35 ± 0.38	1.000	0	0/0
Interp.	67.29 ± 0.61	0.988 ± 0.001	$+0.06 \pm 0.01$	11.8 ± 0.6 / 4.5 ± 0.2
GU (Ours)	74.41 ± 0.77	0.991 ± 0.001	$+0.30 \pm 0.02$	34.0 ± 1.3 / 5.1 ± 0.3

three independent runs using the same missing-image test inputs (TI-T). Linear CKA measures the similarity between fusion-layer representations of each edited model and the original multimodal model under the same TI-T condition. Error-margin gain is measured relative to the original TI-T model. The correction and regression rates denote the fractions of originally incorrect predictions corrected and originally correct predictions broken after editing, respectively.

Both edits preserve the broad representation structure, with linear CKA of 0.988 for interpolation and 0.991 for GU, consistent with their similar UMAP layouts in Figure 4(a). Despite this global similarity, GU produces substantially larger task-level changes: its error-margin gain is $+0.30$ versus $+0.06$, and it corrects 34.0% of the original errors versus 11.8% for interpolation, while increasing the regression rate by only 0.6 percentage points (5.1% versus 4.5%). Correspondingly, TI-T F1 reaches 74.41 with GU versus 67.29 with interpolation. These results show that GU preserves overall representation structure while making targeted changes that improve task-relevant margins and correct substantially more errors.

F.3 More results on robustness–preservation trade-off

The main text reports family-level averages for robustness recovery and full-modality preservation on SNLI-VE. Table 19 provides complete per-variant results, reported as mean $\pm$ sample standard deviation over three independent runs. For each model variant, TI-T and TI-TI denote missing-modality and full-modality performance after editing, while ΔT and ΔTI measure their changes relative to the original TI model.

Across variants, GU generally yields larger TI-T gains than layer interpolation while producing smaller or comparable TI-TI changes. The difference is most pronounced in cross-attention fusion, where GU substantially improves missing-modality robustness with limited full-modality degradation. For CLIP-style and BLIP-2 architectures, the gains are smaller, consistent with their more distributed cross-modal interactions. Overall, the per-variant results reinforce the family-level robustness–preservation pattern reported in Section 6.3.

F.4 Comparisons with more baselines

F.4.1 MoRA and OGM-GE

We further compare GU with two architecture-compatible baselines. MoRA [73] is a parameter-efficient missing-modality adaptation method and is evaluated on CLIP Late Concat and Mean Pooling. Its formulation couples modality-specific and shared low-rank updates across paired text and vision encoder weights. Applying it to Fusion Concat would require new insertion points and a fusion-specific sharing rule, so we retain its matched dual-encoder parameterization.

Table 19: Per-variant robustness–preservation results on SNLI-VE. GU generally provides stronger recovery with comparable or better full-modality preservation than interpolation.

Family	Variant	Target Layer	Method	TI-T	TI-TI	ΔT	ΔTI
Fusion	Concat	fusion_mlp0	Interp.	$74.86 \pm .68$	$86.10 \pm .16$	$1.82 \pm .48$	$-0.12 \pm .10$
	Concat	fusion_mlp0	**GU**	$\mathbf{75.11} \pm .56$	$\mathbf{86.14} \pm .12$	$\mathbf{2.07} \pm .32$	$\mathbf{-0.08} \pm .05$
	Gate	gate_mlp0	Interp.	$71.88 \pm .47$	$86.40 \pm .18$	$0.18 \pm .24$	$-0.13 \pm .11$
	Gate	gate_mlp0	**GU**	$\mathbf{72.05} \pm .35$	$\mathbf{86.44} \pm .12$	$\mathbf{0.35} \pm .07$	$\mathbf{-0.09} \pm .04$
	Cross-Attn	attn_v←t_in	Interp.	$67.29 \pm .61$	$86.50 \pm .24$	$0.94 \pm .29$	$-0.14 \pm .15$
	Cross-Attn	attn_v←t_in	**GU**	$\mathbf{74.41} \pm .77$	$\mathbf{86.53} \pm .17$	$\mathbf{8.06} \pm .40$	$\mathbf{-0.11} \pm .08$
Fusion (strong)	Concat	fusion_mlp0	Interp.	$72.93 \pm .33$	$76.62 \pm .18$	$0.15 \pm .16$	$0.18 \pm .13$
	Concat	fusion_mlp0	**GU**	$\mathbf{73.06} \pm .24$	$\mathbf{76.74} \pm .12$	$\mathbf{0.28} \pm .05$	$\mathbf{0.30} \pm .07$
	Gate	gate_mlp0	Interp.	$70.13 \pm .52$	$73.09 \pm .23$	$2.48 \pm .34$	$0.37 \pm .17$
	Gate	gate_mlp0	**GU**	$\mathbf{71.99} \pm .70$	$\mathbf{73.30} \pm .16$	$\mathbf{4.34} \pm .46$	$\mathbf{0.58} \pm .09$
	Cross-Attn	attn_v←t_in	Interp.	$73.95 \pm .48$	$75.76 \pm .82$	$0.26 \pm .24$	$-3.95 \pm .75$
	Cross-Attn	attn_v←t_in	**GU**	$\mathbf{74.13} \pm .35$	$\mathbf{77.81} \pm .65$	$\mathbf{0.44} \pm .09$	$\mathbf{-1.90} \pm .57$
Two-tower	Late Concat	fusion_mlp0	Interp.	$73.72 \pm .42$	$77.61 \pm .22$	$0.95 \pm .24$	$-0.44 \pm .17$
	Late Concat	fusion_mlp0	**GU**	$\mathbf{73.78} \pm .51$	$\mathbf{77.74} \pm .15$	$\mathbf{1.01} \pm .30$	$\mathbf{-0.31} \pm .09$
	Mean Pooling	text_proj1	Interp.	$75.28 \pm .67$	$74.65 \pm .59$	$1.13 \pm .52$	$-1.97 \pm .55$
	Mean Pooling	text_proj1	**GU**	$\mathbf{75.58} \pm .53$	$\mathbf{75.65} \pm .46$	$\mathbf{1.43} \pm .35$	$\mathbf{-0.97} \pm .41$
VLM	BLIP-2	q_crossattn0_o	Interp.	$70.65 \pm .79$	$81.72 \pm .30$	$0.01 \pm .10$	$0.00 \pm .13$
	BLIP-2	q_crossattn0_o	**GU**	$\mathbf{70.65} \pm .68$	$\mathbf{81.73} \pm .18$	$\mathbf{0.01} \pm .04$	$\mathbf{0.01} \pm .04$

OGM-GE [37] is a training-time modality-balancing method and is evaluated on Fusion Concat and both CLIP variants. CLIP Mean Pooling permits exact decomposition of the full logit into modality contributions. For nonlinear Late Concat, we retain the published discrepancy, gradient-modulation, and generalization-enhancement rules while estimating modality confidence through native unimodal forward paths.

Table 20 reports TI-T / TI-TI as mean $\pm$ sample standard deviation over three independent runs. The ΔTI rows denote GU's mean TI-TI change relative to Null-token. GU achieves the highest mean TI-T in all nine settings, including all six CLIP settings. Across the six CLIP comparisons, its average advantages over MoRA and OGM-GE are 4.42 and 5.04 F1, respectively. Recovery over Null-token is larger on IU-XRay and SNLI-VE. On HM Fusion Concat, the original missing-modality gap is only 0.50 F1, leaving little room for further recovery.

Across all nine settings, GU improves TI-T by 0.94 F1 on average while changing TI-TI by only -0.41 F1; the largest TI-TI loss is 1.08 F1 on IU-XRay Fusion Concat. MoRA and OGM-GE do not consistently improve missing-image performance, whereas GU remains effective across both CLIP variants beyond the original fusion-only comparison.

F.4.2 LoRA and Dual CE

We compare GU with post-hoc LoRA and Dual CE on SNLI-VE using Fusion Cross-Attention (Fusion CA), CLIP Mean Pooling (CLIP MP), and BLIP-2. Table 21 reports performance under missing-image (TI-T) and full-modality (TI-TI) evaluation.

Comparison with LoRA. LoRA starts from the same trained TI checkpoint as GU and adapts the same target layer with rank 8. It trains on missing-image inputs for 4 epochs at learning rate 10^{-4}, selecting the best checkpoint on the development set. GU instead uses the matched unimodal reference for a single closed-form edit, without further gradient updates to the TI checkpoint.

GU achieves higher mean TI-T and TI-TI than LoRA on Fusion CA and CLIP MP. On BLIP-2, LoRA obtains slightly greater TI-T recovery but reduces TI-TI, whereas GU leaves both metrics essentially unchanged. Averaged across the three architectures, LoRA improves TI-T by 2.44 F1 with a TI-TI change of -1.99 F1 relative to Null-token; GU improves TI-T by 3.17 F1 with a TI-TI change

Table 20: Cross-architecture comparison with MoRA and OGM-GE. GU achieves the highest mean TI-T in all nine evaluated settings. Entries report TI-T / TI-TI F1. ΔTI denotes GU's mean TI-TI change relative to Null-token (GU minus Null-token).

Method	Fusion Concat	CLIP Late Concat	CLIP Mean Pooling
Hateful Memes			
Null-token	68.96 ± 0.44 / 69.46 ± 0.32	66.58 ± 0.41 / 67.31 ± 0.27	66.87 ± 0.24 / 67.14 ± 0.20
MoRA	– / –	65.64 ± 0.28 / 65.39 ± 0.83	60.45 ± 6.17 / 64.98 ± 0.87
OGM-GE	66.71 ± 0.56 / 67.07 ± 0.78	66.07 ± 0.47 / 65.85 ± 0.41	64.51 ± 0.66 / 65.68 ± 1.10
GU	69.01 ± 0.24 / 69.32 ± 0.30	67.03 ± 0.17 / 67.26 ± 0.31	66.96 ± 0.34 / 67.09 ± 0.22
GU ΔTI	-0.14	-0.05	-0.05
IU-XRay			
Null-token	80.70 ± 0.38 / 86.41 ± 0.22	79.45 ± 0.29 / 80.85 ± 0.19	74.30 ± 0.38 / 78.00 ± 0.20
MoRA	– / –	73.87 ± 1.91 / 78.14 ± 0.26	70.55 ± 2.06 / 75.13 ± 0.45
OGM-GE	72.62 ± 3.87 / 77.53 ± 2.25	75.04 ± 0.48 / 75.51 ± 1.02	67.54 ± 0.89 / 71.36 ± 0.33
GU	81.69 ± 0.72 / 85.33 ± 0.80	80.16 ± 0.24 / 80.57 ± 0.31	75.98 ± 0.61 / 77.28 ± 0.36
GU ΔTI	-1.08	-0.28	-0.72
SNLI-VE			
Null-token	73.04 ± 0.24 / 86.22 ± 0.07	72.77 ± 0.21 / 78.05 ± 0.06	74.15 ± 0.18 / 76.62 ± 0.06
MoRA	– / –	70.28 ± 0.99 / 75.06 ± 0.19	72.21 ± 0.12 / 72.17 ± 0.04
OGM-GE	67.26 ± 0.33 / 83.91 ± 0.24	68.10 ± 2.14 / 71.07 ± 0.10	68.00 ± 0.54 / 70.30 ± 0.21
GU	75.11 ± 0.56 / 86.14 ± 0.12	73.78 ± 0.51 / 77.74 ± 0.15	75.58 ± 0.53 / 75.65 ± 0.46
GU ΔTI	-0.08	-0.31	-0.97

Table 21: Comparison with post-hoc LoRA and Dual CE on SNLI-VE. Entries report F1 as mean ± sample standard deviation over three independent runs.

Architecture	Method	TI-T	TI-TI
Fusion CA	Null-token	66.35 ± 0.38	86.64 ± 0.10
	LoRA	72.37 ± 0.25	85.02 ± 0.35
	Dual CE	79.92 ± 0.23	85.70 ± 0.22
	GU	74.41 ± 0.77	86.53 ± 0.17
	Dual CE+GU	80.09 ± 0.30	86.51 ± 0.19
CLIP MP	Null-token	74.15 ± 0.18	76.62 ± 0.06
	LoRA	75.18 ± 0.07	72.96 ± 0.64
	Dual CE	76.31 ± 0.13	75.22 ± 0.08
	GU	75.58 ± 0.53	75.65 ± 0.46
	Dual CE+GU	75.46 ± 0.61	75.45 ± 0.78
BLIP-2	Null-token	70.64 ± 0.70	81.72 ± 0.18
	LoRA	70.90 ± 0.86	81.04 ± 0.23
	Dual CE	75.05 ± 0.13	81.19 ± 0.16
	GU	70.65 ± 0.68	81.73 ± 0.18
	Dual CE+GU	75.05 ± 0.17	81.04 ± 0.49

of only -0.36 F1. Thus, GU provides 0.73 F1 greater missing-image recovery with 1.63 F1 less full-modality degradation on average. These results demonstrate a stronger average cross-architecture robustness–preservation trade-off than matched-layer LoRA adaptation, extending the comparison beyond layer interpolation and the fusion setting.

Comparison with Dual CE. Dual CE retrains each architecture from scratch with a shared classifier jointly supervising full-input and missing-image paths, using an auxiliary weight of 0.5. Dual CE+GU applies GU to the resulting Dual CE checkpoint. Dual CE is a strong retraining baseline, improving TI-T over Null-token by 13.57, 2.16, and 4.41 F1 on Fusion CA, CLIP MP, and BLIP-2, respectively. This recovery comes with TI-TI changes of -0.94, -1.40, and -0.53 F1, whereas GU preserves higher full-modality performance on all three architectures.

Table 22: Target-layer selection on IU-XRay Fusion Cross-Attention. Joint angle–harm selection yields the strongest recovery and preservation.

	Angle only	Harm only	**Angle+harm**
Selected layer	`attn_t_from_v_out_proj`	`text_proj`	`attn_v_from_t_in_proj`
TI-T	80.34 ± 0.66	80.72 ± 0.40	$\mathbf{81.41} \pm 0.63$
TI-TI	83.96 ± 0.37	84.21 ± 0.32	$\mathbf{84.44} \pm 0.28$
ΔT	$+0.34$	$+0.72$	$\mathbf{+1.41}$
ΔTI	-0.61	-0.36	$\mathbf{-0.13}$

Table 23: Controlled comparison of editing strategies on Fusion Cross-Attention. GU achieves the highest mean TI-T and TI-TI across all three datasets. Entries report TI-T / TI-TI F1.

Dataset	Weight	Linear	**GU (Ours)**
HM	58.50 ± 0.91 / 66.52 ± 0.33	58.07 ± 1.03 / 66.62 ± 0.21	$\mathbf{60.88} \pm 0.73$ / $\mathbf{67.55} \pm 0.26$
IX	80.14 ± 0.49 / 82.67 ± 0.56	79.83 ± 0.75 / 82.39 ± 0.36	$\mathbf{81.41} \pm 0.63$ / $\mathbf{84.44} \pm 0.28$
SV	67.29 ± 0.61 / 86.50 ± 0.24	66.84 ± 1.29 / 86.50 ± 0.07	$\mathbf{74.41} \pm 0.77$ / $\mathbf{86.53} \pm 0.17$

The two approaches are complementary in the Fusion setting: applying GU after Dual CE raises TI-TI from 85.70 to 86.51 while slightly improving TI-T from 79.92 to 80.09. On CLIP and BLIP-2, the composition provides no additional recovery benefit. Overall, Dual CE favors stronger missing-image recovery through retraining, while GU provides a post-training correction with stronger full-modality preservation. The computational costs of GU, LoRA, and Dual CE are compared in Table 4.

F.5 Controlled interventions for geometric correction

We examine two design choices underlying GU: *where to edit and how to edit*. These controlled comparisons connect the diagnostics in Section 4 with the geometric rationale in Appendix D.2.

F.5.1 Target-layer selection

Using the IU-XRay Fusion Cross-Attention layer sweep in Figure 4(b), we compare angle-only, harm-only, and joint angle–harm selection while keeping checkpoints and the editing protocol fixed. Angle-only and harm-only rank candidate layers by principal-subspace deviation and task-aware harm, respectively, while the joint rule uses the development-set score in Eq. (1). Table 22 reports mean $\pm$ sample standard deviation over three independent runs. ΔT and ΔTI denote changes in TI-T and TI-TI relative to the original TI model.

The joint rule produces the largest TI-T gain ($+1.41$ F1) and smallest TI-TI loss (-0.13 F1). Angle-only and harm-only selection yield smaller gains of $+0.34$ and $+0.72$, with TI-TI changes of -0.61 and -0.36, respectively. Combining geometric deviation with task-aware harm therefore identifies a more effective correction target than either signal alone.

F.5.2 Editing strategy comparison

We next compare three editing strategies on Fusion Cross-Attention across the three datasets. All methods use matched TI/T checkpoints, the same target layer and reference, and the same evaluation protocol. **Weight** interpolates the full selected weight matrix with its reference using fixed $\alpha = 0.5$, thereby changing input orientation together with output factors, scales, and residual. **Linear** and **GU** instead edit the aligned principal input subspace while retaining the original TI output factor, principal coefficients, and residual. Linear applies polar-retracted interpolation as in Eq. (13), whereas GU follows the Grassmannian geodesic. Both use $k = 8$ and $\eta = 0.75$. Table 23 reports TI-T / TI-TI F1 as mean $\pm$ sample standard deviation over three independent runs.

GU achieves the highest mean TI-T and TI-TI in all three settings. On HM and IU-XRay, it improves both missing-image recovery and full-modality preservation over Weight and Linear. On SNLI-VE, GU reaches 74.41 TI-T F1 versus 67.29 for Weight and 66.84 for Linear, while maintaining comparable TI-TI performance. In particular, the comparison with Linear isolates the interpolation

Table 24: A-OKVQA multiple-choice accuracy (%) over ten runs. GU provides the strongest full-modality preservation among the evaluated interventions. Bold marks the best result among Interpolation, LoRA, and GU.

Method	TI-T Acc.	TI-TI Acc.
T Refer. (T-T)	47.17 ± 0.31	–
Shuffle-I	–	39.85 ± 1.97
Orig. TI	38.94 ± 1.92	50.79 ± 0.71
Interp.	39.80 ± 1.21	43.21 ± 1.44
LoRA	$\mathbf{43.52} \pm 0.60$	45.01 ± 1.72
GU (Ours)	41.07 ± 1.79	$\mathbf{47.40} \pm 1.19$

Table 25: Missing-text evaluation with Fusion Cross-Attention. GU improves TI-I while largely preserving full-modality performance across all three datasets.

Method	Eval.	HM	IU-XRay	SNLI-VE
Orig. TI	TI-TI	68.82 ± 0.48	84.57 ± 0.28	86.64 ± 0.10
I Refer.	I-I	66.58 ± 0.15	75.32 ± 0.41	68.74 ± 0.00
Orig. TI	TI-I	63.32 ± 0.88	73.58 ± 0.85	64.48 ± 0.97
Interp.	TI-I	64.10 ± 0.24	73.70 ± 0.76	65.87 ± 0.46
Interp.	TI-TI	68.30 ± 0.28	84.39 ± 0.36	86.01 ± 0.15
GU (Ours)	TI-I	64.34 ± 0.20	74.42 ± 0.22	66.61 ± 0.27
GU (Ours)	TI-TI	68.65 ± 0.23	84.64 ± 0.48	86.67 ± 0.15

path: even with the same target subspace, rank, coefficient, and retained components, the geodesic yields substantially stronger recovery.

Together, these controls support the two key design choices in GU: combining geometric deviation with task-aware harm for target selection, and correcting the selected principal input subspace along a Grassmannian geodesic.

F.6 Generalization to multiple-choice VQA

To extend the evaluation beyond binary classification, we evaluate GU on the A-OKVQA [42] multiple-choice VQA benchmark. We use the strong-encoder cross-attention fusion configuration described in Appendix E.4.1, with Qwen2.5-3B and OpenCLIP ViT-bigG-14 encoders, and follow the official dataset splits.

Table 24 reports accuracy as mean $\pm$ sample standard deviation over ten independent runs. TI-T and TI-TI denote missing-image and full-modality evaluation, respectively; the T reference reports T-T accuracy in the TI-T column. Shuffle-I evaluates the original TI model with shuffled images under full-input inference. Bold marks the best result among Interpolation, LoRA, and GU.

The ordering TI-TI $>$ T-T $>$ TI-T persists on A-OKVQA ($50.79 > 47.17 > 38.94$). Shuffling images reduces full-input accuracy from 50.79% to 39.85%, confirming that image information contributes substantially to the full-input prediction.

GU improves TI-T accuracy by 2.13 points and achieves the highest TI-TI accuracy among the three interventions, with a 3.39-point loss relative to the original TI model. Its TI-TI accuracy exceeds Interpolation and LoRA by 4.19 and 2.39 points, respectively. LoRA achieves larger TI-T recovery ($+4.58$), but with a larger TI-TI loss (-5.78). These results extend GU's recovery–preservation behavior beyond binary classification to multiple-choice VQA.

F.7 Complementary evaluation under missing-text inputs

We further evaluate GU under missing-text inputs using Fusion Cross-Attention on HM, IU-XRay, and SNLI-VE. Here, TI-I denotes a TI-trained model evaluated without text, while I-I denotes the

matched image-only reference. Both GU and layer interpolation use this image-only model as the reference. Table 25 reports F1 as mean $\pm$ sample standard deviation over three independent runs.

The ordering TI-TI $>$ I-I $>$ TI-I holds on all three datasets, mirroring the missing-image pattern in Section 4.2. GU improves TI-I by 1.02, 0.84, and 2.13 F1 on HM, IU-XRay, and SNLI-VE, respectively, while changing TI-TI by only -0.17, $+0.07$, and $+0.03$ F1.

GU also achieves higher mean TI-I than matched layer interpolation and remains closer to the original TI-TI performance on all three datasets. Although the edited TI-I remains below the matched I-I reference, GU consistently narrows this gap. These results demonstrate GU's generalizability across missingness directions, as the same correction principle consistently improves robustness for both missing-image and missing-text inputs when paired with the modality available at deployment.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are supported by Sections 4, 5, 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: In Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In Section 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We use publicly available datasets, as described in Section 6 and Appendix E. We are preparing the code for public release.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix F.1 reports sample standard deviations over three independent runs and ten-seed paired sign-flip tests with Holm correction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Section 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research fully conforms with NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Appendix B.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Properly cited and stated.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research in this paper does not involve crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research in this paper does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development does not involve LLMs as an important, original, or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.