No Free Alignment: Observability-Aware Alignment for Multimodal Heterogeneous Learning

Canran Xiao[1] **Puning Zhao**[1] **Enneng Yang**[1]
Xiaochun Cao[1] **Haobo Fu**[2] **Li Shen**[1*]
[1]Shenzhen Campus of Sun Yat-sen University, China
[2]Tencent AI Platform Department, China
{xiaocr3,zhaopn,yangenn,caoxiaochun,shenli6}@mail.sysu.edu.cn
haobofu@tencent.com

Abstract

Multimodal heterogeneous learning aims to train reliable multimodal models when data sources differ in modality availability, semantic coverage, domain distribution, and missingness patterns. Although cross-modal alignment is widely used to mitigate such heterogeneity, we argue that alignment is not free: enforcing relations that are weakly supported by the observed data can amplify noise, over-transfer missing-modality information, and induce negative transfer. This paper studies when a modality–modality–semantic relation is sufficiently observable to be safely aligned and shared across heterogeneous sources. We propose **ObsAlign**, an observability-aware framework that summarizes visible modality–semantic evidence through compact source-level sketches and estimates relation-level support from bridge evidence, side support, cross-modal consistency, and prototype uncertainty. The resulting observability score gates prototype alignment, calibrates weak missing-modality transfer, and guides evidence-aware source consolidation so that reliable relations are shared while unsupported variation remains local. Across image–text, audio–visual–text, action-recognition, sensor, and healthcare benchmarks under source-partitioned multimodal heterogeneity, ObsAlign achieves the best overall performance in both encoder-based and VLM-compatible regimes, while sharply reducing negative transfer on low-observability relations. These results suggest that relation-level observability provides a practical principle for deciding when cross-modal alignment should be strengthened, weakened, or avoided. Code is publicly available at https://github.com/Xiaocanran1230/ObsAlign.

1 Introduction

Multimodal heterogeneous learning (MHL) trains multimodal models across sources with different modality availability, domain distributions, semantic coverage, supervision granularity, and missingness mechanisms [Nguyen et al., 2025, Phung et al., 2025]. Such heterogeneous observations arise in healthcare, mobile sensing, affective computing, video understanding, and vision–language systems, where each source may provide a partial and biased view of the multimodal world [López et al., 2026, Yuan et al., 2026, Dai et al., 2025]. Conventional multimodal learning often assumes complete, well-aligned modality pairs [Zhao et al., 2026a]. MHL must assess the available cross-modal evidence before deciding how modalities should interact.

A common strategy aligns heterogeneous modalities through shared embeddings and contrastive objectives [Chen et al., 2026], prototype matching [Qian et al., 2026], or distillation, prompt adaptation, and missing-modality completion [Zhao et al., 2026b]. While such alignment can transfer

*Corresponding author: Li Shen.

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

useful information across modalities, it also introduces a hidden assumption: the relation being aligned is sufficiently supported by the observed data. Under heterogeneous observations, this assumption can fail. Two modalities may be frequently observed in isolation but rarely co-occur for a particular semantic concept; a relation may be reliable in one source but spurious in another; and a missing-modality proxy may become overconfident when the target relation is not identifiable. Thus, cross-modal alignment is not free: unsupported alignment can create negative transfer.

We study alignment at the level of modality–modality–semantic relations. For example, image–text evidence may be abundant for one semantic class but almost absent for another. Whether image and text should be aligned therefore depends on relation-level evidence, including bridge observations, side support, cross-source consistency, and uncertainty. We call this property *semantic observability*: a relation is observable if the heterogeneous data sources provide enough reliable evidence to identify and transfer it.

We first conduct a diagnostic study showing that alignment utility is strongly relation-dependent. Uniform alignment improves well-observed relations but frequently harms weakly observed ones. Relations with similar modality frequency can exhibit opposite transfer effects depending on their bridge evidence. This observation motivates an observability-aware alignment mechanism.

Our contributions are threefold:

- We identify *relation-level observability* as a missing principle in multimodal heterogeneous learning, showing that marginal modality availability is insufficient to decide whether cross-modal alignment is beneficial.
- We propose **ObsAlign**, an observability-aware alignment framework that estimates modality–modality–semantic support from compact source sketches, gates alignment by relation observability, calibrates weak missing-modality transfer, and consolidates source evidence according to bridge informativeness.
- Across diverse MHL benchmarks spanning image–text, audio–visual–text, sensor, action-recognition, and healthcare tasks, ObsAlign improves over strong heterogeneous multimodal baselines and reduces negative transfer on low-observability relations.

2 Related Work

Multimodal heterogeneous learning. Multimodal learning represents, aligns, and fuses heterogeneous signals [Baltrušaitis et al., 2018]. We define multimodal heterogeneous learning (MHL) as learning from sources that vary in modality availability, semantic coverage, domain distribution, missingness, or resources, subsuming missing-modality learning [Wu et al., 2026], source-partitioned multimodal learning [Feng et al., 2023], and multimodal federated learning [Che et al., 2023, Lin et al., 2023, Pan et al., 2024]. While existing methods address heterogeneity at the sample, modality, client/source, or global-representation level, we focus on modality–modality–semantic relations and ask whether each relation is observable enough to justify alignment.

Missing-modality robust multimodal learning. Missing-modality learning seeks robust prediction when some modalities are absent. Existing methods compensate for missing inputs through parameter-efficient adaptation or missing-case-aware prompts [Reza et al., 2024, Hu et al., 2024], predict or complete absent-modality representations [Kim and Kim, 2024, Ke et al., 2025], and distill knowledge across modalities [Wang et al., 2023, Li et al., 2024]. In source-partitioned settings, prior work further explores prototype masking or transfer, reconstruction, feature imputation, cross-modal aggregation, contrastive regularization, and reconfigurable representations [Bao et al., 2023, Wang et al., 2024, Nguyen et al., 2024, Poudel et al., 2025, Liu et al., 2025, Le et al., 2025, Nguyen et al., 2025]. These methods recover or transfer missing information. ObsAlign asks whether the target cross-modal relation is observable enough to make that transfer reliable.

Cross-modal alignment and negative transfer. Contrastive learning formalizes alignment and uniformity [Wang and Isola, 2020], while CLIP, ViLT, ALBEF, and ImageBind learn shared or fused representations across image–text and broader modality spaces [Radford et al., 2021, Kim et al., 2021, Li et al., 2021a, Girdhar et al., 2023]. Reversible primitive–composition alignment preserves compositional structure during continual vision–language adaptation [Xiao et al., 2026]. Modality-gap analyses show that contrastive spaces may retain systematic modality separation [Liang

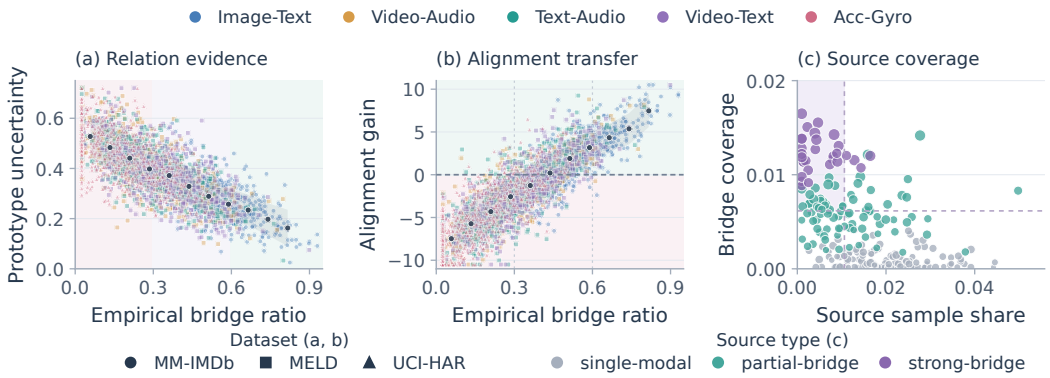

Figure 1: **Exploratory diagnosis of no-free alignment.** Relations in (a)–(b) are colored by modality pair and shaped by dataset (dark dots: binned means; bands: ± 1 SD); (c) plots sources. **(a)** Weak bridge evidence accompanies higher prototype uncertainty. **(b)** Uniform alignment improves well-bridged relations but often harms weakly bridged ones. **(c)** High-bridge sources can have small sample shares, so sample-weighted consolidation can underweight relation-critical sources.

et al., 2022], and negative-transfer studies show that weakly related or poorly supported sources can degrade performance [Zhang et al., 2022].

Distributed and federated multimodal heterogeneity. Federated methods harmonize non-IID data with diffusion models [Chen et al., 2024b] and share prototype-anchored prompts for continual personalization [Xiao and Hou, 2026]. Multimodal federated learning also faces incomplete modalities and communication constraints. Multimodal methods use split architectures, cross-modal activity recognition, modality-aware training, contrastive ensembles, selective feature alignment, multimodal transformers, hypergraph aggregation, and modality/client selection [Chen and Zhang, 2022, Yang et al., 2024, Ouyang et al., 2023, Yu et al., 2023, Tan et al., 2023, Sun et al., 2024, Qi and Li, 2024, Yuan et al., 2024, 2026]. Recent foundation-model variants further adapt pretrained multimodal models via heterogeneous finetuning, MLLM assistance, instruction or prompt tuning, privacy-preserving learning, and ownership protection [Chen et al., 2024a, Zhang et al., 2025, Xiong et al., 2025, Singha et al., 2025, Phung et al., 2025, Tran et al., 2025, Xiao et al., 2025, Gai et al., 2025]. ObsAlign complements these approaches with relation-level semantic observability to decide which cross-modal evidence should be shared, weakened, or kept local.

3 Preliminary Diagnosis: When Does Alignment Help?

Question. Cross-modal alignment is often used to connect heterogeneous sources in a shared multimodal space. However, under multimodal heterogeneity, the evidence supporting a relation can be highly uneven: two modalities may be frequent in isolation but rarely co-observed for a particular semantic unit. We therefore ask a diagnostic question: *when does alignment help, and when does it cause negative transfer?*

Diagnostic setup. On MM-IMDb, MELD, and UCI-HAR, two strategies share source partitions, modality masks, backbone, and optimization budget. *Task-only SourceAvg* optimizes the task objective alone. *Uniform Alignment* applies equal alignment strength to all locally supported modality–modality–semantic relations. For each relation $r = (m, n, s)$, we measure its bridge evidence $B_{m,n,s}$, prototype uncertainty $U_{m,n,s}$, and relation-conditioned alignment gain $G_{m,n,s}$. We also compare each source's sample share q_i with its bridge coverage R_i to diagnose whether sample-weighted consolidation underweights structurally important bridge sources. The exact definitions, relation filtering rules, validation protocol, and visualization details are provided in Appendix B.1.

Findings. Figure 1 shows three consistent patterns. First, relation support is highly non-uniform: low-bridge relations tend to have larger prototype uncertainty, indicating that marginal modality availability alone does not make a relation identifiable. Second, uniform alignment has a sign-changing effect: it benefits sufficiently bridged relations but frequently produces negative transfer when bridge evidence

is weak. Third, bridge coverage is not proportional to sample share, which suggests that standard sample-weighted consolidation can suppress sources that are small but structurally important for cross-modal identification.

Implication. These observations motivate a relation-level treatment of multimodal alignment. Reliable relations should receive stronger alignment, weakly supported relations should receive weaker alignment, and bridge sources should have sufficient influence during consolidation. ObsAlign implements this principle by estimating relation-level observability and using it to gate alignment and guide evidence-aware cross-source consolidation.

4 Method

We propose **ObsAlign**, an observability-aware alignment framework for multimodal heterogeneous learning. ObsAlign operates on a collection of heterogeneous sources, each of which may expose different modality subsets, semantic distributions, domains, and missingness patterns. ObsAlign estimates the observability of each modality–modality–semantic relation and uses it to regulate alignment, missing-modality transfer, and cross-source consolidation. At each coordination round, a coordinator broadcasts the shared model, semantic prototypes, and relation-level observability scores. Each source performs source-specific learning with observability-gated alignment, computes compact semantic sketches from its visible modalities, and uploads only shared-parameter updates and sketches. The coordinator then refreshes the observability map and consolidates source updates with evidence-aware weights. This procedure determines which relations are aligned, how much each source influences the shared model, and which variation remains source-specific. Figure 2 shows the pipeline.

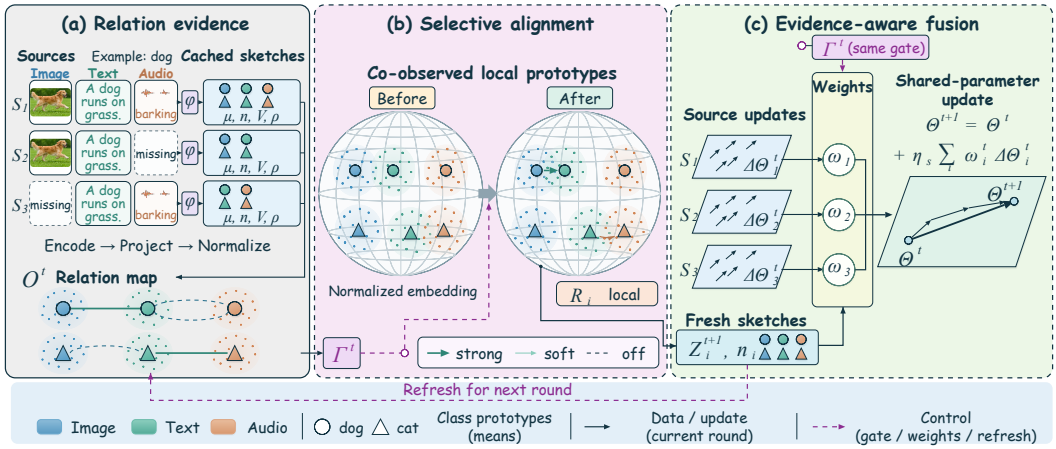

Figure 2: **Pipeline of ObsAlign.** A coordinator maintains shared modality encoders, semantic prototypes, and relation-level observability scores. Each heterogeneous source performs local/source-specific learning on its observed modalities, aligns only well-supported modality–semantic relations, keeps source-specific residuals private, and summarizes visible modality–semantic evidence through compact semantic sketches. The coordinator then performs evidence-aware cross-source consolidation and refreshes the observability map for the next coordination round.

4.1 Problem Formulation

Let $\mathcal{C} = \{1, \ldots, K\}$ denote the heterogeneous sources, $\mathcal{M} = \{1, \ldots, M\}$ the modalities, and $\mathcal{S}$ the task-conditioned semantic units. ObsAlign models relations $(m, n, s) \in \mathcal{M} \times \mathcal{M} \times \mathcal{S}$. Source i has modality set $\mathcal{M}_i \subseteq \mathcal{M}$. Let $v_{ij,m} \in \{0, 1\}$ indicate whether sample j actually observes modality m, with $v_{ij,m} = 0$ for $m \notin \mathcal{M}_i$. Its observed set is $\mathcal{M}_{ij} = \{m \in \mathcal{M}_i : v_{ij,m} = 1\} \neq \varnothing$, and its data are

$$\mathcal{D}_i = \left\{ \left(\{x_{ij}^{(m)}\}_{m \in \mathcal{M}_{ij}}, y_{ij} \right) \right\}_{j=1}^{n_i}, \qquad a_{ij} = q_\omega \left(\{x_{ij}^{(m)}\}_{m \in \mathcal{M}_{ij}}, y_{ij} \right) \in \Delta^{S-1}. \tag{1}$$

Here, $a_{ij,s}$ is the membership of sample j to semantic unit s, and q_ω is fixed during source-partitioned training. In supervised tasks, a_{ij} is the class indicator. In unlabeled tasks, q_ω is a frozen semantic assigner.

For modality m, ObsAlign uses a shared encoder $f_m(\cdot;\theta_m)$, a projection head $P_m(\cdot;\phi_m)$, and a source-specific residual branch $r_{i,m}(\cdot;R_{i,m})$. For observed inputs $m \in \mathcal{M}_{ij}$, we define

$$\begin{aligned}
\tilde{z}_{ij}^{(m)} &= \mathrm{Norm}\big(P_m(f_m(x_{ij}^{(m)};\theta_m);\phi_m)\big),\\
z_{ij}^{(m)} &= \mathrm{Norm}\big(P_m(f_m(x_{ij}^{(m)};\theta_m) + r_{i,m}(x_{ij}^{(m)};R_{i,m});\phi_m)\big),\\
\mathcal{L}_i^{\mathrm{task}} &= \frac{1}{n_i}\sum_{j=1}^{n_i} \ell\left(h\left(\frac{1}{|\mathcal{M}_{ij}|}\sum_{m\in\mathcal{M}_{ij}} z_{ij}^{(m)};\psi\right), y_{ij}\right).
\end{aligned} \tag{2}$$

where $\mathrm{Norm}(u) = u/(\|u\|_2+\varepsilon)$, $\tilde{z}_{ij}^{(m)}$ is the shared-branch representation used for coordinator-visible sketches, $z_{ij}^{(m)}$ is the source-adapted representation used for local prediction, $h(\cdot;\psi)$ is the prediction head, and ℓ is the supervised task loss.

4.2 Local Semantic Sketching

Each source summarizes its observed modality–semantic evidence through compact prototypes, support counts, dispersion statistics, and reliability scores.

After local/source-specific training in round t, source i computes a semantic sketch from the shared-branch features $\tilde{z}_{ij}^{(m)}$:

$$n_{i,m,s}^{t+1} = \sum_{j=1}^{n_i} a_{ij,s} v_{ij,m}, \qquad \tilde{\mu}_{i,m,s}^{t+1} = \frac{\sum_{j=1}^{n_i} a_{ij,s} v_{ij,m} \tilde{z}_{ij}^{(m)}}{n_{i,m,s}^{t+1}+\varepsilon}, \tag{3}$$

$$V_{i,m,s}^{t+1} = \frac{\sum_{j=1}^{n_i} a_{ij,s} v_{ij,m} \|\tilde{z}_{ij}^{(m)} - \tilde{\mu}_{i,m,s}^{t+1}\|_2^2}{n_{i,m,s}^{t+1}+\varepsilon}, \qquad \rho_{i,m,s}^{t+1} = \frac{n_{i,m,s}^{t+1}}{n_{i,m,s}^{t+1}+n_0}\exp\left(-\frac{V_{i,m,s}^{t+1}}{\tau_v}\right). \tag{4}$$

where $n_{i,m,s}^{t+1}$ is the observed semantic support, $\tilde{\mu}_{i,m,s}^{t+1} \in \mathbb{R}^d$ is the shared prototype, $V_{i,m,s}^{t+1}$ is its dispersion, and $\rho_{i,m,s}^{t+1} \in [0,1]$ is its reliability. Masked sums include only $v_{ij,m}=1$ terms; absent inputs are not encoded. Zero-support entries have $\rho_{i,m,s}^{t+1}=0$ and are excluded from prototype-based losses. Source i uploads

$$\mathcal{Z}_i^{t+1} = \{n_{i,m,s}^{t+1}, \tilde{\mu}_{i,m,s}^{t+1}, V_{i,m,s}^{t+1}, \rho_{i,m,s}^{t+1}\}_{m\in\mathcal{M}_i, s\in\mathcal{S}}$$

together with the shared-parameter update $\Delta\Theta_i^t$. The source-specific residual R_i remains local.

4.3 Semantic Observability Estimation

A relation between modality m and modality n for semantic unit s should be aligned only when the heterogeneous source collection provides enough evidence to identify that relation. ObsAlign quantifies this identifiability-like property through a relation-level observability score $O_{m,n,s} \in [0,1]$. For the $P = \binom{M}{2}$ unordered modality pairs, the matrix representation is

$$\mathbf{O} \in [0,1]^{S\times P}, \qquad \mathbf{O}_{s,p} := O_{m,n,s}, \qquad p=(m,n),\ m<n.$$

Rows index semantic units and columns index modality pairs.

The coordinator first aggregates source prototypes into shared semantic prototypes:

$$\bar{\mu}_{m,s}^{t+1} = \frac{\sum_{i\in\mathcal{C}_t} e_{i,m,s}^{t+1} n_{i,m,s}^{t+1} \rho_{i,m,s}^{t+1} \tilde{\mu}_{i,m,s}^{t+1}}{\sum_{i\in\mathcal{C}_t} e_{i,m,s}^{t+1} n_{i,m,s}^{t+1} \rho_{i,m,s}^{t+1} + \varepsilon}, \tag{5}$$

$$\begin{aligned}
O_{m,n,s}^{t+1} = \ & \left(1 - e^{-\mathrm{Bridge}_{m,n,s}^{t+1}}\right)\\
& \cdot \sigma\big(\alpha\,\mathrm{Bridge}_{m,n,s}^{t+1} + \beta\,\mathrm{Cons}_{m,n,s}^{t+1} + \eta\,\mathrm{Supp}_{m,n,s}^{t+1} - \delta\,\mathrm{Unc}_{m,n,s}^{t+1} - b_0\big).
\end{aligned} \tag{6}$$

where $e^{t+1}_{i,m,s} \in \{0,1\}$ indicates whether source i provides reliable evidence for (m,s). The term $\mathrm{Supp}_{m,n,s}$ measures side support for both modalities, $\mathrm{Bridge}_{m,n,s}$ measures bridge-source evidence, $\mathrm{Cons}_{m,n,s}$ measures cross-modal consistency, and $\mathrm{Unc}_{m,n,s}$ measures prototype uncertainty. The prefactor $1 - e^{-\mathrm{Bridge}_{m,n,s}}$ measures the sufficiency of bridge evidence. Their exact definitions are given in Appendix A.2. The modality symmetry is $O_{m,n,s} = O_{n,m,s}$. The score is used in the next coordination round. We fix $(\alpha, \beta, \eta, \delta, b_0) = (1, 1, 0.5, 1, 0)$ across datasets. A warm-start sketch pass initializes O^0, and a small validation grid selects the alignment weight, gate interval, and consolidation strength. Appendix C.7 reports the coefficient study and default recipe.

4.4 Observability-Gated Local Learning

Unconditional alignment treats all modality relations as equally transferable. Under heterogeneity, this can force unsupported or source-specific relations into a shared space. ObsAlign converts observability into an alignment gate so that reliable relations are tightened, uncertain relations are softly regularized, and unobservable relations receive no global alignment force.

At the beginning of round t, each source receives the already-computed score O^t. The relation gate is

$$\Gamma^t_{m,n,s} = O^t_{m,n,s} \cdot \mathrm{clip}\left(\frac{O^t_{m,n,s} - \tau_l}{\tau_h - \tau_l}, 0, 1\right). \tag{7}$$

Here, $0 \le \tau_l < \tau_h \le 1$, and $\mathrm{clip}(u, 0, 1)$ truncates u into $[0,1]$. Thus, relations below τ_l are not aligned, relations between τ_l and τ_h are softly aligned, and relations above τ_h are aligned with strength close to $O^t_{m,n,s}$.

Let $\mathcal{P}_{i,s} = \{(m,n) : m,n \in \mathcal{M}_i,\ m < n,\ n_{i,m,s} > 0,\ n_{i,n,s} > 0\}$ contain pairs with observed semantic support on both sides. The prototype-level alignment loss is

$$\mathcal{L}^{\mathrm{align},t}_i = \sum_{s=1}^{S} \sum_{(m,n)\in\mathcal{P}_{i,s}} \Gamma^t_{m,n,s} \left\| \tilde{\mu}_{i,m,s}(\Theta_i) - \tilde{\mu}_{i,n,s}(\Theta_i) \right\|_2^2. \tag{8}$$

Here, $\tilde{\mu}_{i,m,s}(\Theta_i)$ is the current shared-branch prototype during round t, computed using the same observation mask. Each modality prototype uses its own observed samples; alignment operates at the semantic-prototype level.

The extension to missing-modality prototype transfer uses transport maps and is given in Appendix A.4.

The complete source-specific objective is

$$\mathcal{J}^t_i = \mathcal{L}^{\mathrm{task}}_i + \lambda_{\mathrm{align}} \mathcal{L}^{\mathrm{align},t}_i + \lambda_{\mathrm{cal}} \mathcal{L}^{\mathrm{cal},t}_i + \lambda_R \sum_{m\in\mathcal{M}_i} \|R_{i,m}\|_F^2. \tag{9}$$

The calibration term $\mathcal{L}^{\mathrm{cal},t}_i$ penalizes overconfident missing-modality predictions under weak observability and is defined in Appendix A.3. The residual penalty keeps source-specific branches small, so local residuals absorb unsupported source variation without dominating the shared representation.

4.5 Evidence-Aware Cross-Source Consolidation

Sample-weighted consolidation can underweight small but structurally important sources. In MHL, a source with few samples may still provide rare co-observed modality pairs that make a relation observable. ObsAlign therefore uses relation-level bridge evidence to decide how much each source update should influence the shared model.

Let $b^{t+1}_{i,m,n,s} \ge 0$ denote the bridge evidence supplied by source i for relation (m,n,s), as defined in Appendix A.2. ObsAlign computes the source bridge contribution and consolidation weight as

$$\chi^t_i = \sum_{s=1}^{S} \sum_{\substack{m,n\in\mathcal{M}_i \\ m<n}} \Gamma^t_{m,n,s} b^{t+1}_{i,m,n,s}, \qquad \omega^t_i = \frac{n_i(1 + \zeta\chi^t_i)}{\sum_{\ell\in\mathcal{C}_t} n_\ell(1 + \zeta\chi^t_\ell) + \varepsilon}. \tag{10}$$

where χ^t_i is large when source i provides bridge evidence for relations that are already observable at the beginning of round t, and $\zeta \ge 0$ controls the strength of evidence-aware weighting.

Table 1: **Main results under source-partitioned multimodal heterogeneity on image–text and foundation-model-compatible benchmarks.** Feature-based methods use the same extracted modality features or lightweight encoders, while foundation-model methods use a matched VLM backbone and comparable trainable-parameter budget. The best and second-best results are shown in **bold** and underlined, respectively. Our methods are shaded in light purple.

Method	UPMC Acc.	MM-IMDb Macro-F1	Hateful AUROC	CrisisMMD Macro-F1	Avg. Score
Task and source heterogeneity baselines *with the same feature/encoder backbone*					
Local-only	77.42	26.88	61.73	69.54	58.89
Task-only SourceAvg [McMahan et al., 2017]	80.16	28.41	63.25	70.83	60.66
FedProx [Li et al., 2020]	80.73	28.95	63.88	71.26	61.21
FedBN [Li et al., 2021b]	81.24	29.18	64.05	71.58	61.51
FedProto [Tan et al., 2022]	81.68	29.46	64.37	72.05	61.89
Feature-based MHL and missing-modality robust baselines					
FedMSplit [Chen and Zhang, 2022]	82.91	29.91	65.72	72.84	62.85
CreamFL [Yu et al., 2023]	83.36	30.44	66.51	73.36	63.42
FedDAT [Chen et al., 2024a]	84.12	30.87	67.24	73.91	64.04
FedMAC [Nguyen et al., 2024]	84.36	31.05	67.69	74.13	64.31
HAMFL [Qi and Li, 2024]	85.07	31.46	68.29	74.52	64.84
PEPSY [Nguyen et al., 2025]	84.82	31.73	68.54	74.38	64.87
Foundation-model and prompt-adapter baselines *with matched VLM backbone and trainable budget*					
MLLM-LLaVA-FL [Zhang et al., 2025]	85.46	31.82	68.73	74.96	65.24
Pilot [Xiong et al., 2025]	85.88	32.24	69.18	75.34	65.66
FED-PRIME [Phung et al., 2025]	86.37	32.91	69.84	75.81	66.23
FedMVP [Singha et al., 2025]	<u>86.92</u>	<u>33.37</u>	<u>70.36</u>	<u>76.18</u>	<u>66.71</u>
ObsAlign variants *under the corresponding MHL regime*					
OBSALIGN-ENC	86.51	33.08	70.02	75.93	66.39
OBSALIGN-VLM	**88.46**	**35.21**	**72.14**	**77.63**	**68.36**

The coordinator update is

$$\Theta^{t+1} = \Theta^t + \eta_s \sum_{i \in \mathcal{C}_t} \omega_i^t \Delta\Theta_i^t, \qquad \Theta = \{\theta_m, \phi_m\}_{m=1}^M \cup \{\psi\}. \tag{11}$$

Only shared parameters are consolidated. Source-specific residuals remain local. A modality-specific consolidation variant is provided in Appendix A.5.

5 Experiments

5.1 Experimental Setup

We evaluate ObsAlign under source-partitioned MHL, where sources differ in modality availability, semantic coverage, domain distribution, and missingness patterns. This protocol covers missing-modality learning, heterogeneous source adaptation, and federated/distributed multimodal training as special cases.

Benchmarks and protocols. We use a diverse benchmark suite spanning image–text, audio–visual–text, sensor, action-recognition, and healthcare scenarios, including UPMC Food-101, MM-IMDb, Hateful Memes, CrisisMMD, MELD, CREMA-D, UCF101, Moments-in-Time, EPIC-Kitchens, UCI-HAR, KU-HAR, PTB-XL, and Sleep-EDF. Semantic units are instantiated as task labels unless otherwise specified. Natural source identities and modality missingness are separate protocol choices. We use user-, subject-, speaker-, patient-, or participant-level partitions where available and construct Dirichlet/class-skewed sources otherwise. Both groups use controlled modality masks shared by all methods. Dataset statistics, preprocessing, source construction, and missing-modality protocols are detailed in Appendix B.2.

Metrics and evaluation. We follow the standard metric of each dataset family, using accuracy, macro-F1, UAR, or AUROC as appropriate. To evaluate the proposed observability mechanism beyond average performance, we additionally report relation-conditioned gains, negative-transfer rates, severe-missing robustness, calibration, and cost–performance tradeoffs in the corresponding analyses. All results are averaged over three random seeds.

Table 2: **Main results under source-partitioned multimodal heterogeneity on affective, action-recognition, sensor, and healthcare benchmarks.** These benchmarks follow the encoder-based MHL regime. The best and second-best results are shown in **bold** and underlined, respectively. Our method is shaded in light purple.

Method	MELD UAR	CREMA-D UAR	UCF101 Acc.	MiT10 Acc.	EPIC Acc.	UCI-HAR F1	KU-HAR F1	PTB-XL Acc.	Sleep-EDF Acc.
Task and source heterogeneity baselines *with the same feature/encoder backbone*									
Local-only	55.26	60.38	71.62	40.37	29.81	87.28	76.15	55.42	46.38
Task-only SourceAvg [McMahan et al., 2017]	57.14	62.27	73.08	42.26	31.44	89.34	78.27	57.88	47.52
FedProx [Li et al., 2020]	57.72	62.81	73.65	42.93	32.07	89.87	78.96	58.55	48.01
FedBN [Li et al., 2021b]	58.07	63.14	73.96	43.25	32.26	90.43	79.64	58.31	47.76
FedProto [Tan et al., 2022]	58.34	63.43	74.26	43.49	32.58	90.81	80.07	59.04	48.15
Feature-based MHL and missing-modality robust baselines									
FedMSplit [Chen and Zhang, 2022]	59.16	64.18	75.02	44.13	33.06	91.24	80.66	60.13	49.02
CreamFL [Yu et al., 2023]	60.07	65.02	75.63	44.88	33.72	91.78	81.03	61.24	49.63
FedDAT [Chen et al., 2024a]	60.64	65.68	76.28	45.37	34.26	92.16	81.47	62.08	50.04
FedMAC [Nguyen et al., 2024]	60.73	65.77	76.41	45.52	34.43	92.08	81.37	63.87	50.42
HAMFL [Qi and Li, 2024]	60.92	65.95	77.04	46.16	34.91	93.42	82.59	62.71	49.81
PEPSY [Nguyen et al., 2025]	61.38	66.41	76.52	45.78	34.72	92.65	82.08	68.26	54.84
ObsAlign variant *under the encoder-based MHL regime*									
OBSALIGN-ENC	**63.84**	**68.62**	**79.33**	**48.37**	**37.08**	**94.81**	**84.26**	**71.92**	**57.31**

Baselines. We compare against four groups of methods: (i) task/source heterogeneity baselines, including Local-only, Task-only SourceAvg, FedProx, FedBN, and FedProto; (ii) feature-based MHL and missing-modality robust methods, including FedMSplit, CreamFL, FedDAT, HAMFL, FedMAC, and PEPSY; (iii) foundation-model and prompt-adapter baselines, including MLLM-LLaVA-FL, Pilot, FED-PRIME, and FedMVP; and (iv) ObsAlign variants, namely OBSALIGN-ENC and OBSALIGN-VLM. Within each comparison regime, all methods use the same source partitions, modality masks, communication rounds, local training budget, backbone, and trainable-parameter budget whenever applicable. Implementation details are provided in Appendix B.2.

5.2 Main Results

Overall comparison. Tables 1 and 2 report the main comparisons under source-partitioned MHL, with methods grouped by compatible backbone and trainable-component regimes. ObsAlign consistently improves across both settings: OBSALIGN-VLM outperforms the strongest foundation-model baseline FedMVP by 1.65 average points, while OBSALIGN-ENC surpasses feature-based MHL and missing-modality baselines. The improvements also generalize beyond image–text benchmarks to affective computing, action recognition, sensor, and healthcare tasks, with the largest gains on PTB-XL, Sleep-EDF, EPIC-Kitchens, and MiT10, where modality evidence is highly uneven or source-dependent. These results support conditioning cross-modal alignment on relation-level observability.

5.3 Ablation and Mechanism Analysis

Single-factor ablation. We ablate each component of OBSALIGN-ENC in Table 3. The observability gate is most critical, with its removal dropping the overall score from 68.34 to 64.87 and the severe-missing score from 60.72 to 55.96. Semantic-level observability and bridge-aware consolidation each contribute 1.91 points, while the remaining components provide smaller but consistent gains, especially under severe missingness.

Relation-level alignment utility. We examine whether the learned observability score $O_{m,n,s}$ predicts when cross-modal alignment is helpful. For each modality–semantic relation, we compute its relation-conditioned gain over Task-only SourceAvg and visualize this gain against $O_{m,n,s}$ in Figure 3. Uniform Alignment exhibits a sign-changing behavior: low-observability relations are frequently harmed, while high-observability relations benefit from alignment. ObsAlign suppresses the negative-transfer region, reducing the low-observability negative-transfer rate from 91.7% to 3.4%, while increasing the high-observability mean gain from $+3.23$ to $+4.17$. These results support using $O_{m,n,s}$ to select relations for alignment.

Table 3: **Single-factor ablation of ObsAlign.** Each ablation removes or simplifies one component. Scores are percentage points (higher is better), and gray subscripts show drops from the full model. Best and second-best scores are **bold** and underlined. The full method is shaded in light purple.

Variant	Image–Text Avg.	Affective Avg.	Action Avg.	Sensor Avg.	Healthcare Avg.	Severe Avg.	Overall Avg.
OBSALIGN-ENC full	**66.39**	**66.23**	**54.93**	**89.54**	**64.62**	**60.72**	**68.34**
w/o semantic-level observability	64.38$_{\downarrow 2.01}$	64.31$_{\downarrow 1.92}$	52.96$_{\downarrow 1.97}$	87.94$_{\downarrow 1.60}$	62.57$_{\downarrow 2.05}$	57.81$_{\downarrow 2.91}$	66.43$_{\downarrow 1.91}$
w/o observability gate	63.15$_{\downarrow 3.24}$	63.44$_{\downarrow 2.79}$	51.32$_{\downarrow 3.61}$	86.11$_{\downarrow 3.43}$	60.32$_{\downarrow 4.30}$	55.96$_{\downarrow 4.76}$	64.87$_{\downarrow 3.47}$
w/o bridge-aware consolidation	64.62$_{\downarrow 1.77}$	64.79$_{\downarrow 1.44}$	52.87$_{\downarrow 2.06}$	87.62$_{\downarrow 1.92}$	62.24$_{\downarrow 2.38}$	57.88$_{\downarrow 2.84}$	66.43$_{\downarrow 1.91}$
w/o reliability-weighted sketch	65.28$_{\downarrow 1.11}$	65.18$_{\downarrow 1.05}$	53.91$_{\downarrow 1.02}$	88.32$_{\downarrow 1.22}$	63.14$_{\downarrow 1.48}$	58.91$_{\downarrow 1.81}$	67.17$_{\downarrow 1.17}$
w/o uncertainty penalty	65.44$_{\downarrow 0.95}$	65.36$_{\downarrow 0.87}$	54.06$_{\downarrow 0.87}$	88.67$_{\downarrow 0.87}$	63.42$_{\downarrow 1.20}$	59.24$_{\downarrow 1.48}$	67.39$_{\downarrow 0.95}$
w/o local residual branch	65.11$_{\downarrow 1.28}$	64.96$_{\downarrow 1.27}$	53.52$_{\downarrow 1.41}$	88.01$_{\downarrow 1.53}$	62.89$_{\downarrow 1.73}$	58.43$_{\downarrow 2.29}$	66.90$_{\downarrow 1.44}$
w/o calibration loss	65.97$_{\downarrow 0.42}$	65.68$_{\downarrow 0.55}$	54.38$_{\downarrow 0.55}$	89.01$_{\downarrow 0.53}$	63.53$_{\downarrow 1.09}$	58.98$_{\downarrow 1.74}$	67.71$_{\downarrow 0.63}$
w/o dynamic sketch refresh	64.94$_{\downarrow 1.45}$	64.66$_{\downarrow 1.57}$	53.20$_{\downarrow 1.73}$	87.84$_{\downarrow 1.70}$	62.62$_{\downarrow 2.00}$	57.67$_{\downarrow 3.05}$	66.65$_{\downarrow 1.69}$

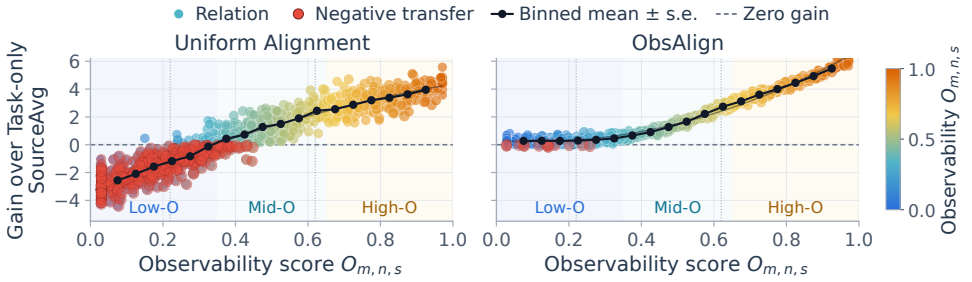

Figure 3: **Relation-level alignment utility under multimodal heterogeneity.**

Relation-separable representation geometry. We further inspect whether observability-gated alignment changes the geometry of modality–semantic prototypes. Figure 4 projects the learned prototypes into two dimensions and connects highly observable cross-modal relations. Compared with Uniform Alignment, ObsAlign yields a more relation-separable geometry: highly observable relations become more compact across modalities, while weakly observable relations are not forcibly collapsed into incorrect shared regions. Quantitatively, ObsAlign reduces the average cross-modal prototype distance for high-observability relations from 0.84 to 0.41, while increasing the inter-semantic margin from 1.03 to 1.42. This geometry supports selective tightening of reliable relations while preserving semantic separation.

5.4 Relation Selection and Practical Evaluation

Pre-alignment ordering. To separate the ordering signal from the final model, we compute O_r^0 from a common Task-only warm-start checkpoint and then align each relation alone for the same short update budget. Across MM-IMDb, MELD, and PTB-XL, pre-intervention O^0 attains mean Spearman $\rho = 0.75$ and AUROC 0.91 for positive intervention gain, compared with 0.57 and 0.81 for bridge-only ordering.

Alignment-budget control. Table 4 matches total alignment mass across methods. Top-K and hard-gate methods also match ObsAlign's 54.2% active-relation coverage, while Uniform-Budget retains all relations. ObsAlign exceeds Uniform-Budget by 3.42 points and Bridge-TopK by 1.62 points. On FedSEA's compatible subset, it improves the average from 65.84 to 67.58. The compatible-subset comparison is detailed in Appendix C.5.

Misleading bridges and deployment. With 20% semantic-mismatch or domain-shortcut corruption and unchanged support/co-observation counts, ObsAlign's false-high observability rates are 14.7%/11.6%, compared with 52.4%/57.4% for Bridge-only (Appendix C.6). The sketch adds a mean 0.24% to the model-update upload, with a maximum of 0.98%, and server refresh averages 4.7 ms per round. Sketches can reveal semantic distributions. Log-binned counts and clipped, noisy prototypes reduce an empirical distribution-inference attack's R^2 from 0.71 to 0.23, at a 0.76-point overall cost.

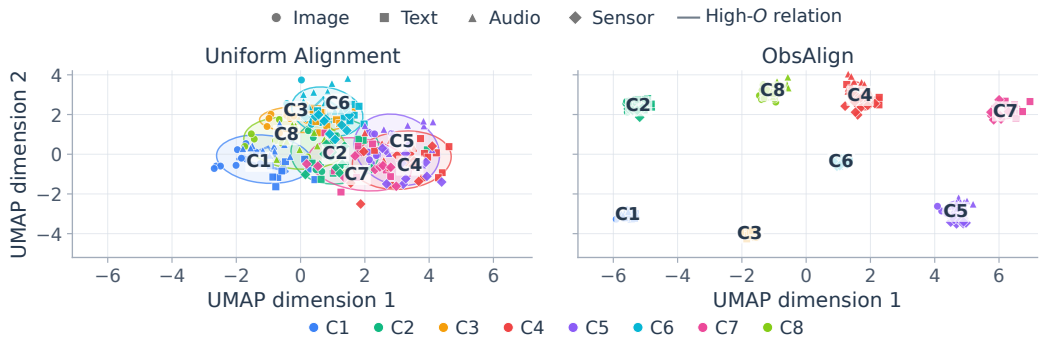

Figure 4: **Relation-separable geometry under observability-aware alignment.**

Table 4: Relation selection at matched total alignment mass. Overall and Severe use the encoder-based benchmark-family aggregation of Table 3. Low-O negative-transfer rate (NTR) follows the relation diagnostic protocol.

Method	Overall ↑	Severe ↑	Low-O NTR ↓
Uniform-Budget	64.92	55.26	72.5%
Bridge-TopK	66.72	58.30	14.9%
ObsAlign-Hard	67.61	59.72	7.6%
ObsAlign	**68.34**	**60.72**	**3.4%**

6 Conclusion

This paper studies when cross-modal alignment should be applied in multimodal heterogeneous learning, where modality evidence is often incomplete, unevenly distributed, and source-dependent. We propose **ObsAlign**, an observability-aware framework that estimates relation-level support from compact source sketches and uses it to regulate three key decisions: which modality–semantic relations should be aligned, how missing-modality transfer should be calibrated, and how heterogeneous source updates should be consolidated. Across diverse MHL benchmarks, ObsAlign consistently improves overall performance and reduces negative transfer on weakly observable relations where uniform alignment is unreliable. These results suggest that relation-level observability provides a practical and interpretable principle for multimodal heterogeneous learning: reliable evidence should be shared, uncertain relations should be weakened, and unsupported source-specific variation should remain local.

Limitations. Observability depends on the chosen semantic units and on counterevidence visible within or across sources. Spurious relations shared consistently by all sources remain a failure case for observational sketches. The benchmarks use controlled modality masks, and sketch exchange can expose source-level semantic statistics. Formal differential privacy requires an additional mechanism.

Acknowledgments and Disclosure of Funding

Funding. This work is supported by NSFC Grant (No. 62576364, 62625602), Guang-Dong Basic and Applied Basic Research Foundation (2026B1515020071), Shenzhen Science and Technology Program (NO. JCYJ20241202124430041), Shenzhen Science and Technology Program(NO.SYSRD20250529113401002), Shenzhen Science and Technology Program (NO. ZDCY20250901113005006), China Postdoctoral Science Foundation (2026M791721), Postdoctoral Fellowship Program of CPSF (GZC20260899), CCF-Tencent Rhino-Bird Open Research Fund (NO. CCF-Tencent RAGR20250114), Tencent JR2025TEG002, CAAI-Tencent Rhino-Bird Open Research Fund (NO. 2026CAAI-Tencent-08).

Competing interests. The authors declare no competing interests relevant to this work.

References

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

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra Perez, and Jorge Luis Reyes Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21st International European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 437–442, 2013.

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

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

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv preprint arXiv:2312.13508*, 2023.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23(15):6986, 2023.

Haokun Chen, Yao Zhang, Denis Krompass, Jindong Gu, and Volker Tresp. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11285–11293, 2024a.

Jiajun Chen, Sai Cheng, Yutao Yuan, Yirui Zhang, Haitao Yuan, Peng Peng, and Yi Zhong. PROMISE: Prompt-attentive hierarchical contrastive learning for robust cross-modal representation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20076–20082, 2026. doi: 10.1609/aaai.v40i24.39093.

Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 87–96, 2022.

Xiaohong Chen, Canran Xiao, and Yongmei Liu. Confusion-resistant federated learning via diffusion-based data harmonization on Non-IID data. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-4368.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24507–24517, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4035–4045, 2023.

Keke Gai, Dongjue Wang, Jing Yu, Mohan Wang, Liehuang Zhu, and Qi Wu. Mfl-owner: ownership protection for multi-modal federated learning via orthogonal transform watermark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3049–3058, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

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.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25864–25873, 2025.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

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.

Donggeun Kim and Taesup Kim. Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In *European Conference on Computer Vision*, pages 171–187. Springer, 2024.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.

Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, Eui-Nam Huh, and Choong Seon Hong. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion*, 122:103219, 2025.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021a.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021b.

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

Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, 20(4): 539–553, 2023.

Junming Liu, Guosun Zeng, Ding Wang, Yanting Gao, and Yufei Jin. FedRecon: Missing modality reconstruction in distributed heterogeneous environments. *arXiv preprint arXiv:2504.09941v1*, 2025. URL `https://arxiv.org/abs/2504.09941v1`.

Alberto López, John Zobolas, Tanguy Dumontier, and Tero Aittokallio. Multi-modal learning with incomplete data. *Nature Communications*, 17(1):10356, 2026. doi: 10.1038/s41467-026-77212-w.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Manh Duong Nguyen, Trung Thanh Nguyen, Huy Hieu Pham, Trong Nghia Hoang, Phi Le Nguyen, and Thanh Trung Huynh. Fedmac: Tackling partial-modality missing in federated learning with cross-modal aggregation and contrastive regularization. In *2024 22nd International Symposium on Network Computing and Applications (NCA)*, pages 278–285. IEEE, 2024.

Manh Duong Nguyen, Trong Nghia Hoang, Thanh Trung Huynh, Quoc Viet Hung Nguyen, and Phi Le Nguyen. Learning reconfigurable representations for multimodal federated learning with missing data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pages 530–543, 2023.

Hao Pan, Xiaoli Zhao, Lipeng He, Yicong Shi, and Xiaogang Lin. A survey of multimodal federated learning: background, applications, and perspectives. *Multimedia Systems*, 30(4):222, 2024.

Thu Hang Phung, Duong M Nguyen, Thanh Trung Huynh, Quoc Viet Hung Nguyen, Trong Nghia Hoang, and Phi Le Nguyen. Federated prompt-tuning with heterogeneous and incomplete multimodal client data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3936–3946, 2025.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536, 2019.

Pranav Poudel, Aavash Chhetri, Prashnna Gyawali, Georgios Leontidis, and Binod Bhattarai. Multimodal federated learning with missing modalities through feature imputation network. In *Annual Conference on Medical Image Understanding and Analysis*, pages 289–299. Springer, 2025.

Fan Qi and Shuai Li. Adaptive hyper-graph aggregation for modality-agnostic federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12312–12321, 2024.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. DecAlign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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.

Niloy Sikder and Abdullah-Al Nahid. Ku-har: An open dataset for heterogeneous human activity recognition. *Pattern Recognition Letters*, 146:46–54, 2021.

Mainak Singha, Subhankar Roy, Sarthak Mehrotra, Ankit Jha, Moloud Abdar, Biplab Banerjee, and Elisa Ricci. Fedmvp: Federated multimodal visual prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17869–17878, 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards multi-modal transformers in federated learning. In *European Conference on Computer Vision*, pages 229–246. Springer, 2024.

Min Tan, Yinfu Feng, Lingqiang Chu, Jingcheng Shi, Rong Xiao, Haihong Tang, and Jun Yu. Fedsea: Federated learning via selective feature alignment for non-iid multimodal data. *IEEE Transactions on Multimedia*, 26:5807–5822, 2023.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 8432–8440, 2022.

Linh Tran, Wei Sun, Stacy Patterson, and Ana Milanova. Privacy-preserving personalized federated prompt learning for multimodal large language models. *arXiv preprint arXiv:2501.13904*, 2025.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

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.

Shu Wang, Zhe Qu, Yuan Liu, Shichao Kan, Yixiong Liang, and Jianxin Wang. Fedmmr: Multi-modal federated learning via missing modality reconstruction. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2015.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Canran Xiao and Liwei Hou. Prototype-aligned federated soft-prompts for continual web personalization. In *Proceedings of the ACM Web Conference 2026*, 2026. doi: 10.1145/3774904.3792626.

Canran Xiao, Tianxiang Xu, Siyuan Ma, Yiyang Jiang, Haoyu Gao, and Yuhan Wu. Reversible primitive–composition alignment for continual vision–language learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://proceedings.iclr.cc/paper_files/paper/2026/hash/8f0d446441a938d9de420a8ab8d7fd36-Abstract-Conference.html`.

Tianzhe Xiao, Yichen Li, Yining Qi, Yi Liu, Haozhao Wang, Yi Wang, Ruixuan Li, et al. Enhancing privacy in multimodal federated learning with information theory. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Pilot: Building the federated multimodal instruction tuning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21716–21724, 2025.

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5345–5361, 2024.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations*, 2023.

Liangqi Yuan, Dong-Jun Han, Vishnu Pandi Chellapandi, Stanislaw H Zak, and Christopher G Brinton. Fedmfs: Federated multimodal fusion learning with selective modality communication. In *ICC 2024-IEEE International Conference on Communications*, pages 287–292. IEEE, 2024.

Liangqi Yuan, Dong-Jun Han, Su Wang, Devesh Upadhyay, and Christopher G Brinton. Communication-efficient multimodal federated learning: Joint modality and client selection. *IEEE Transactions on Mobile Computing*, 2026.

Jianyi Zhang, Hao Yang, Ang Li, Xin Guo, Pu Wang, Haiming Wang, Yiran Chen, and Hai Li. Mllm-llava-fl: Multimodal large language model assisted federated learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4066–4076. IEEE, 2025.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

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*, 2026a.

Yunpeng Zhao, Cheng Chen, Qing You Pang, Yibing Fu, Quanzheng Li, Carol Tang, Beng Ti Ang, and Yueming Jin. Tackling dual-stage missing modalities in brain tumor segmentation via robust modality reconstruction and prompt-guided modality adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13314–13322, 2026b. doi: 10.1609/aaai.v40i16.38334.

A Additional Technical Details for ObsAlign

A.1 Semantic Assignments and Client Sketches

ObsAlign uses a fixed semantic assignment function q_ω to map each local sample into a distribution over semantic units. For supervised classification, q_ω is the label indicator. For unlabeled data, q_ω is a frozen clustering model trained before federated optimization. The assignment function remains fixed during ObsAlign training.

The sketch in Eqs. (3)–(4) uses shared-branch features $\tilde{z}_{ij}^{(m)}$ only where $v_{ij,m} = 1$. The same mask is used for the count, prototype, and dispersion. A zero-support entry is assigned zero reliability and omitted from prototype-based losses. This choice keeps private residuals out of server-side semantic prototypes. The uploaded sketch for client i at round $t+1$ is

$$\mathcal{Z}_i^{t+1} = \left\{ n_{i,m,s}^{t+1}, \tilde{\mu}_{i,m,s}^{t+1}, V_{i,m,s}^{t+1}, \rho_{i,m,s}^{t+1} : m \in \mathcal{M}_i,\ s \in \mathcal{S} \right\}. \tag{12}$$

The support term $n_{i,m,s}^{t+1}$ controls whether a local prototype is statistically meaningful, while the dispersion term $V_{i,m,s}^{t+1}$ controls whether the prototype is concentrated. Their product form in $\rho_{i,m,s}^{t+1}$ downweights small and unstable local semantic clusters.

A.2 Observability Components

The binary evidence indicator used in Eq. (5) is

$$e_{i,m,s}^{t+1} = \mathbf{1}[m \in \mathcal{M}_i]\, \mathbf{1}[n_{i,m,s}^{t+1} \geq n_{\min}]\, \mathbf{1}[\rho_{i,m,s}^{t+1} \geq \rho_{\min}]. \tag{13}$$

Here, $n_{\min} > 0$ is the minimum observed support and $\rho_{\min} \in [0, 1]$ is the reliability threshold. A client contributes only when enough samples observe the modality for that semantic unit and its prototype is reliable.

Support. The effective support for modality m and semantic unit s is

$$A_{m,s}^{t+1} = \sum_{i \in \mathcal{C}_t} e_{i,m,s}^{t+1} n_{i,m,s}^{t+1} \rho_{i,m,s}^{t+1}. \tag{14}$$

For relation (m, n, s), ObsAlign uses the weaker side as the support bottleneck:

$$\mathrm{Supp}_{m,n,s}^{t+1} = \min\left\{ \log(1 + A_{m,s}^{t+1}), \log(1 + A_{n,s}^{t+1}) \right\}. \tag{15}$$

This term is high only when both modalities have reliable evidence for the same semantic unit.

Bridge evidence. A client provides a semantic-prototype bridge for (m, n, s) when both modality–semantic nodes have reliable observed support. This source-level evidence is computed from the two masked prototypes. Its bridge evidence is

$$b_{i,m,n,s}^{t+1} = e_{i,m,s}^{t+1} e_{i,n,s}^{t+1} \rho_{i,m,s}^{t+1} \rho_{i,n,s}^{t+1} \exp\left(-\frac{\|\tilde{\mu}_{i,m,s}^{t+1} - \tilde{\mu}_{i,n,s}^{t+1}\|_2^2}{\tau_b} \right). \tag{16}$$

Here, $\tau_b > 0$ is a temperature. The exponential term assigns larger bridge evidence when the two modality prototypes are locally consistent in the shared semantic space.

The direct bridge strength is

$$D_{m,n,s}^{t+1} = \sum_{i \in \mathcal{C}_t} b_{i,m,n,s}^{t+1}. \tag{17}$$

To capture transitive bridge evidence, ObsAlign constructs a semantic-unit-specific modality graph with adjacency

$$H_s^{t+1}(m, n) = D_{m,n,s}^{t+1}, \qquad H_s^{t+1}(m, m) = 0. \tag{18}$$

Let $\widehat{H}_s^{t+1}$ be the degree-normalized adjacency matrix. The multi-hop bridge score is

$$P_{m,n,s}^{t+1} = \left[\sum_{\ell=2}^{L} \gamma_\ell \left(\widehat{H}_s^{t+1} \right)^\ell \right]_{m,n}, \qquad \gamma_\ell = \gamma_0^{\ell-1}, \quad 0 < \gamma_0 < 1. \tag{19}$$

The final bridge term in Eq. (6) is

$$\mathrm{Bridge}_{m,n,s}^{t+1} = \log\left(1 + D_{m,n,s}^{t+1} + \gamma_{\mathrm{path}} P_{m,n,s}^{t+1}\right). \tag{20}$$

The logarithm prevents high-frequency semantic units from dominating the score. The prefactor in Eq. (6) is

$$1 - e^{-\mathrm{Bridge}_{m,n,s}^{t+1}} = \frac{D_{m,n,s}^{t+1} + \gamma_{\mathrm{path}} P_{m,n,s}^{t+1}}{1 + D_{m,n,s}^{t+1} + \gamma_{\mathrm{path}} P_{m,n,s}^{t+1}}.$$

It is zero without direct or multi-hop bridge evidence and approaches one as their combined strength grows.

Consistency. ObsAlign estimates a lightweight semantic transport map using bridge clients:

$$T_{m\to n}^{(s),t+1} = \arg\min_{T\in\mathbb{R}^{d\times d}} \sum_{i\in\mathcal{C}_t} b_{i,m,n,s}^{t+1} \left\| T\tilde{\mu}_{i,m,s}^{t+1} - \tilde{\mu}_{i,n,s}^{t+1} \right\|_2^2 + \lambda_T \|T - I\|_F^2. \tag{21}$$

The regularizer $\lambda_T\|T - I\|_F^2$ stabilizes the map when bridge support is limited.

The consistency score is defined by cycle preservation:

$$\mathrm{Cons}_{m,n,s}^{t+1} = \exp\left(-\frac{\|\bar{\mu}_{m,s}^{t+1} - T_{n\to m}^{(s),t+1} T_{m\to n}^{(s),t+1} \bar{\mu}_{m,s}^{t+1}\|_2^2 + \|\bar{\mu}_{n,s}^{t+1} - T_{m\to n}^{(s),t+1} T_{n\to m}^{(s),t+1} \bar{\mu}_{n,s}^{t+1}\|_2^2}{2\tau_c}\right). \tag{22}$$

When $D_{m,n,s}^{t+1} < d_{\min}$, ObsAlign sets $\mathrm{Cons}_{m,n,s}^{t+1} = 0$. This prevents consistency from being inferred without bridge evidence.

Uncertainty. The modality-side prototype uncertainty is

$$\mathcal{V}_{m,s}^{t+1} = \frac{\sum_{i\in\mathcal{C}_t} e_{i,m,s}^{t+1} n_{i,m,s}^{t+1} \rho_{i,m,s}^{t+1} \left(V_{i,m,s}^{t+1} + \|\tilde{\mu}_{i,m,s}^{t+1} - \bar{\mu}_{m,s}^{t+1}\|_2^2\right)}{A_{m,s}^{t+1} + \varepsilon}. \tag{23}$$

The first term inside the parentheses measures within-client dispersion, and the second measures cross-client disagreement.

The relation-level uncertainty is

$$\mathrm{Unc}_{m,n,s}^{t+1} = \frac{1}{2}\left(\mathcal{V}_{m,s}^{t+1} + \mathcal{V}_{n,s}^{t+1}\right) + \frac{1}{1 + D_{m,n,s}^{t+1}}. \tag{24}$$

The second term increases uncertainty when direct bridge evidence is scarce.

A.3 Calibration under Weak Observability

ObsAlign regularizes weakly observable relations toward calibrated uncertainty in missing-modality predictions. For client i, visible modality m, target modality n, and semantic unit s, define the prototype-induced distribution

$$q_{i,m\to n}^{t}(r \mid s) = \frac{\exp\left(\cos\left(\tilde{\mu}_{i,m,s}(\Theta_i), \bar{\mu}_{n,r}^{t}\right)/\tau_q\right)}{\sum_{r'=1}^{S} \exp\left(\cos\left(\tilde{\mu}_{i,m,s}(\Theta_i), \bar{\mu}_{n,r'}^{t}\right)/\tau_q\right)}. \tag{25}$$

Here, $q_{i,m\to n}^{t}(r \mid s)$ measures which semantic unit r in modality n is predicted from semantic unit s in modality m.

Let $u(r) = 1/S$ be the uniform distribution over semantic units. The calibration loss in Eq. (9) is

$$\mathcal{L}_i^{\mathrm{cal},t} = \sum_{s=1}^{S} \sum_{\substack{m\in\mathcal{M}_i \\ n_{i,m,s}>0}} \sum_{n\notin\mathcal{M}_i} \left(1 - O_{m,n,s}^{t}\right) \mathbf{1}[O_{m,n,s}^{t} < \tau_h] \,\mathrm{KL}\left(q_{i,m\to n}^{t}(\cdot \mid s) \,\middle\|\, u\right). \tag{26}$$

Minimizing this term discourages confident prototype transfer for weakly observable missing modalities.

A.4 Missing-Modality Prototype Transfer

The main objective aligns locally supported semantic prototypes. For transfer to a modality absent from the client, ObsAlign uses the transport map in Eq. (21) to map the missing-modality prototype into the visible modality space. The transferred alignment loss is

$$\mathcal{L}_i^{\mathrm{miss},t} = \sum_{s=1}^{S} \sum_{\substack{m \in \mathcal{M}_i \\ n_{i,m,s}>0}} \sum_{n \notin \mathcal{M}_i} \Gamma^t_{m,n,s} \left\| \tilde{\mu}_{i,m,s}(\Theta_i) - T^{(s),t}_{n\to m} \bar{\mu}^t_{n,s} \right\|_2^2. \tag{27}$$

Mapping the global prototype into the visible modality space accounts for modality-specific geometry during transfer.

When this extension is used, the local objective becomes

$$\mathcal{J}^t_{i,\mathrm{full}} = \mathcal{J}^t_i + \lambda_{\mathrm{miss}} \mathcal{L}_i^{\mathrm{miss},t}. \tag{28}$$

The main experiments use Eq. (9); Eq. (28) defines the optional transfer extension.

A.5 Modality-Specific Bridge-Aware Aggregation

The main text uses a client-level aggregation weight. ObsAlign also supports modality-specific aggregation when the global model has modality-specific encoders. For modality m, the relation contribution of client i is

$$\chi^t_{i,m} = \sum_{s=1}^{S} \sum_{\substack{n \in \mathcal{M}_i \\ n \neq m}} \Gamma^t_{m,n,s} b^{t+1}_{i,m,n,s}. \tag{29}$$

The modality-specific aggregation weight is

$$\omega^t_{i,m} = \frac{n_i \mathbf{1}[m \in \mathcal{M}_i](1+\zeta\chi^t_{i,m})}{\sum_{\ell \in \mathcal{C}_t} n_\ell \mathbf{1}[m \in \mathcal{M}_\ell](1+\zeta\chi^t_{\ell,m}) + \varepsilon}. \tag{30}$$

The modality encoder and projection head are updated by

$$\theta^{t+1}_m = \theta^t_m + \eta_s \sum_{i \in \mathcal{C}_t} \omega^t_{i,m} \Delta\theta^t_{i,m}, \qquad \phi^{t+1}_m = \phi^t_m + \eta_s \sum_{i \in \mathcal{C}_t} \omega^t_{i,m} \Delta\phi^t_{i,m}. \tag{31}$$

For the prediction head, ObsAlign uses the client-level weight in Eq. (10):

$$\psi^{t+1} = \psi^t + \eta_s \sum_{i \in \mathcal{C}_t} \omega^t_i \Delta\psi^t_i. \tag{32}$$

A.6 Computational and Communication Complexity

For client i, the semantic sketch contains one prototype, support count, dispersion value, and reliability value for each visible modality–semantic pair. The communication cost of the sketch is

$$\mathcal{O}(|\mathcal{M}_i| S d), \tag{33}$$

where S is the number of semantic units and d is the prototype dimension. The cost is independent of the local sample count.

On the server, computing direct bridge evidence over selected clients costs

$$\mathcal{O}(|\mathcal{C}_t| M^2 S). \tag{34}$$

Computing multi-hop bridge evidence by dense matrix powers costs

$$\mathcal{O}(SLM^3), \tag{35}$$

where L is the maximum path length. Since multimodal benchmarks usually have small M, the dominant server-side factor is the number of semantic units S.

Algorithm 1 ObsAlign Training under Source-Partitioned MHL

Require: Sources $\mathcal{C}$, modalities $\mathcal{M}$, semantic units $\mathcal{S}$, initial shared model Θ^0.

Warm start: collect initial semantic sketches and compute $\bar{\mu}^0$ and O^0 using Eqs. (3)–(6).
for coordination round $t = 0, 1, \ldots, T-1$ **do**
  Coordinator samples sources $\mathcal{C}_t$ and broadcasts $\Theta^t, \bar{\mu}^t, O^t$.
  for each source $i \in \mathcal{C}_t$ in parallel **do**
    Optimize $\mathcal{J}_i^t$ in Eq. (9) on local/source data $\mathcal{D}_i$.
    Compute semantic sketch $\mathcal{Z}_i^{t+1}$ from shared-branch features using Eqs. (3)–(4).
    Upload $\Delta\Theta_i^t$ and $\mathcal{Z}_i^{t+1}$; keep R_i local.
  end for
  Coordinator computes ω_i^t using Eq. (10) and updates Θ^{t+1} using Eq. (11).
  Coordinator refreshes $\bar{\mu}^{t+1}$ and O^{t+1} using Eqs. (5)–(6).
end for
return Shared model Θ^T and observability tensor O^T; each source retains its local residual R_i^T.

A.7 Initialization

ObsAlign performs a warm-start sketching pass before the first training round. The server broadcasts Θ^0, selected clients compute $\mathcal{Z}_i^0$, and the server initializes $\bar{\mu}^0$ and O^0. The first trainable round then uses O^0 in Eq. (7).

A.8 Scope: Source-Partitioned MHL and Federated Instantiations

ObsAlign is formulated for source-partitioned multimodal heterogeneous learning. A source can represent a user, subject, patient, device, institution, domain, data provider, or any partition that exposes only a partial view of the full multimodal space. The formulation allows sources to differ in modality availability, semantic coverage, domain distribution, data quality, or missingness mechanism.

Federated or distributed training is one concrete instantiation of this source-partitioned setting. In that case, sources correspond to clients, the coordinator corresponds to a server, and source updates are communicated through a federated optimization protocol. The relation-level observability estimator also applies when multimodal evidence is partitioned across heterogeneous sources, including centralized multi-domain training where raw data can be pooled, institutional training where raw data cannot be shared, or hybrid settings where only sketches and model updates are exchanged.

We therefore use the term *source* in the main formulation and reserve *client* for federated instantiations. Similarly, *SourceAvg* denotes sample-weighted consolidation over heterogeneous sources, and reduces to FedAvg when the sources are federated clients.

A.9 Training Procedure

Algorithm 1 summarizes the overall training procedure of ObsAlign. The sketch–observe–align loop starts with warm-start sketches that initialize relation-level observability. Each selected source learns using the map available at the start of its round; newly uploaded sketches refresh the map for the next round.

B Additional Experimental Details

B.1 Details of the Exploratory Relation Diagnostic

The diagnostic in Section 3 isolates a failure mechanism of evidence-agnostic cross-modal alignment under source-partitioned multimodal heterogeneity. We use Image–Text relations from MM-IMDb; Video–Audio, Text–Audio, and Video–Text relations from MELD; and accelerometer–gyroscope (Acc–Gyro) relations from UCI-HAR.

Controlled baselines. We compare two deliberately simple strategies under identical source partitions, modality masks, backbone architecture, communication rounds, local epochs, and optimizer

settings. The first strategy, *Task-only SourceAvg*, optimizes only the task objective:

$$\mathcal{L}_i^{\text{task-only}} = \mathcal{L}_i^{\text{task}}. \tag{36}$$

The second strategy, *Uniform Alignment*, adds a single global alignment strength to all locally supported modality–modality–semantic relations:

$$\mathcal{L}_i^{\text{uniform}} = \mathcal{L}_i^{\text{task}} + \lambda_{\text{uni}} \sum_{s=1}^{S} \sum_{(m,n)\in\mathcal{P}_{i,s}} \|\hat{\mu}_{i,m,s} - \hat{\mu}_{i,n,s}\|_2^2 . \tag{37}$$

Here, $\hat{\mu}_{i,m,s}$ is the mini-batch semantic center of modality m and semantic unit s, using only samples with $v_{ij,m} = 1$. The set $\mathcal{P}_{i,s}$ uses the corresponding mini-batch supports, so only centers with positive observed support are aligned. The same λ_{uni} is used for all relations, sources, and communication rounds. Both strategies use sample-weighted source averaging:

$$\Theta^{t+1} = \Theta^t + \eta_s \sum_{i\in\mathcal{C}_t} \frac{n_i}{\sum_{\ell\in\mathcal{C}_t} n_\ell} \Delta\Theta_i^t. \tag{38}$$

The strategies differ only in the uniform alignment objective; model capacity, source partitioning, and aggregation are held fixed.

Semantic units and diagnostic relation set. We use the same semantic assignment function q_ω as in Eq. (1). For supervised classification, q_ω is the one-hot label indicator. For multi-label or soft-assignment settings, $a_{ij,s} \in [0,1]$ is used as a sample weight when computing relation-level statistics.

We index each modality–modality–semantic relation by $r = (m,n,s)$, where $m < n$ follows a fixed ordering over modalities. To avoid unstable estimates from degenerate relations, we keep only relations with sufficient side support for both modalities and sufficient validation support:

$$\mathcal{R}_{\text{diag}} = \left\{(m,n,s) : N_{m,s}^{\text{tr}} \geq \nu_{\text{side}},\ N_{n,s}^{\text{tr}} \geq \nu_{\text{side}},\ N_{m,n,s}^{\text{val}} \geq \nu_{\text{val}}\right\}. \tag{39}$$

The support terms are

$$N_{m,s}^{\text{tr}} = \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{D}_i^{\text{tr}}} a_{ij,s} v_{ij,m}, \qquad N_{m,n,s}^{\text{val}} = \sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{D}_i^{\text{val}}} a_{ij,s} v_{ij,m} v_{ij,n}. \tag{40}$$

The set $\mathcal{R}_{\text{diag}}$ retains weakly bridged relations so that the diagnostic captures the effect of limited co-observation.

Bridge evidence. For each relation $r = (m,n,s) \in \mathcal{R}_{\text{diag}}$, we measure empirical bridge evidence using the training partition:

$$B_{m,n,s} = \frac{\sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{D}_i^{\text{tr}}} a_{ij,s} v_{ij,m} v_{ij,n}}{\sum_{i\in\mathcal{C}} \sum_{j\in\mathcal{D}_i^{\text{tr}}} a_{ij,s} \max\{v_{ij,m}, v_{ij,n}\} + \varepsilon}. \tag{41}$$

The numerator counts semantic-weighted samples that actually observe both modalities; the denominator counts those observing at least one. Thus, $B_{m,n,s}$ measures sample-level co-observation, distinct from the source-level semantic-prototype bridge used by the observability estimator. Large marginal support can coexist with small $B_{m,n,s}$ when the modalities occur in different samples or sources.

Prototype uncertainty. We compute prototype uncertainty post hoc from the representation learned by Task-only SourceAvg. This measures cross-source disagreement before an explicit alignment loss is injected. For each source, modality, and semantic unit, we compute a held-out diagnostic center:

$$\bar{z}_{i,m,s} = \frac{\sum_{j\in\mathcal{D}_i^{\text{val}}} a_{ij,s} v_{ij,m}\, \tilde{z}_{ij}^{(m)}}{\sum_{j\in\mathcal{D}_i^{\text{val}}} a_{ij,s} v_{ij,m} + \varepsilon}. \tag{42}$$

Let $n_{i,m,s}^{\text{val}} = \sum_{j\in\mathcal{D}_i^{\text{val}}} a_{ij,s} v_{ij,m}$. The mean $\bar{z}_{m,s}$ weights nonempty centers by $n_{i,m,s}^{\text{val}}$. The modality-side and relation-level uncertainty scores are

$$U_{m,s} = \frac{\sum_{i\in\mathcal{C}} n_{i,m,s}^{\text{val}} \|\bar{z}_{i,m,s} - \bar{z}_{m,s}\|_2^2}{\sum_{i\in\mathcal{C}} n_{i,m,s}^{\text{val}} + \varepsilon}, \qquad U_{m,n,s} = \frac{1}{2}\left(U_{m,s} + U_{n,s}\right). \tag{43}$$

Large $U_{m,n,s}$ indicates cross-source disagreement about the position of a semantic unit in the modality-specific representation space. This is a post-hoc diagnostic, separate from both baselines' training objectives.

Relation-conditioned alignment gain. To measure whether uniform alignment helps a specific relation, we evaluate both strategies under the same relation-conditioned validation view. For relation (m, n, s), we use validation examples in which the modality pair (m, n) is available and compute

$$\mathrm{Score}^{a}_{m,n,s} = \mathrm{Metric}\Big(\{(\hat{y}^{a,(m,n)}_{ij}, y_{ij}, a_{ij,s}) : i \in \mathcal{C},\ j \in \mathcal{D}^{\mathrm{val}}_i,\ v_{ij,m}v_{ij,n} = 1\}\Big), \quad a \in \{\mathrm{task}, \mathrm{align}\}. \tag{44}$$

Here, $\hat{y}^{a,(m,n)}_{ij}$ is the prediction of strategy a when evaluated using the same modality-pair view (m, n), and $a_{ij,s}$ restricts the metric to semantic unit s. The metric follows the dataset protocol; for loss-based tasks, we use negative loss so that larger values always indicate better performance. The relation-conditioned gain of uniform alignment is

$$G_{m,n,s} = \mathrm{Score}^{\mathrm{align}}_{m,n,s} - \mathrm{Score}^{\mathrm{task}}_{m,n,s}. \tag{45}$$

A positive $G_{m,n,s}$ means that uniform alignment improves the relation, while a negative value indicates negative transfer.

Source bridge coverage. To diagnose consolidation bias, we compare each source's sample share with its relation bridge coverage. The sample share is

$$q_i = \frac{n_i}{\sum_{\ell \in \mathcal{C}} n_\ell}. \tag{46}$$

The bridge coverage of source i is computed over the diagnostic relation set:

$$R_i = \frac{1}{|\mathcal{R}_{\mathrm{diag}}|} \sum_{(m,n,s) \in \mathcal{R}_{\mathrm{diag}}} \frac{\sum_{j \in \mathcal{D}^{\mathrm{tr}}_i} a_{ij,s} v_{ij,m} v_{ij,n}}{\sum_{\ell \in \mathcal{C}} \sum_{j \in \mathcal{D}^{\mathrm{tr}}_\ell} a_{\ell j,s} v_{\ell j,m} v_{\ell j,n} + \varepsilon}. \tag{47}$$

The contrast between q_i and R_i reveals whether sources that dominate sample-weighted consolidation are also the sources that provide relation-level bridge evidence. A source can have small q_i but large R_i when it contains rare co-observed modality pairs for important semantic units.

Visualization protocol. Panels (a)–(b) of Fig. 1 show a round-wise longitudinal diagnostic. Each plotted observation is indexed by

$$(d, k, t, m, n, s),$$

where d denotes the dataset, k the random seed, t the communication round, and $(m, n, s) \in \mathcal{R}_{\mathrm{diag}}$ the modality–modality–semantic relation. The same relation can therefore contribute multiple points across rounds and seeds. The definitions above suppress d, k, t for readability; prototype uncertainty and alignment gain use the corresponding round's model states. Panel (a) plots bridge evidence $B_{m,n,s}$ against prototype uncertainty $U_{m,n,s}$. Panel (b) plots bridge evidence $B_{m,n,s}$ against relation-conditioned alignment gain $G_{m,n,s}$. Each point in Panel (c) corresponds to one source and plots sample share q_i against bridge coverage R_i. Observations across seeds and rounds are pooled for visualization; diagnostic quantities are computed separately for each dataset, seed, and round.

Data usage. Bridge evidence $B_{m,n,s}$, source sample share q_i, and source bridge coverage R_i are computed from the training partition. Prototype uncertainty $U_{m,n,s}$ and alignment gain $G_{m,n,s}$ are computed on the validation partition for post-hoc diagnosis. Diagnostic statistics and hyperparameter selection use the training and validation partitions. The diagnostic motivates ObsAlign, whose training uses the sketch-based observability estimator.

B.2 Main Experiment Details

Overview of comparison regimes. Different MMFL methods make different assumptions about modality encoders, trainable components, and pretrained backbones. We organize experiments into two matched regimes. The first is the *encoder-based MMFL regime*, where all methods use the same lightweight modality encoders or the same pre-extracted modality features. This regime is used for generic FL, feature-based MMFL, missing-modality FL, sensor, action-recognition, and healthcare benchmarks. The corresponding variant of our method is denoted as OBSALIGN-ENC. The second is the *foundation-model MMFL regime*, where all methods use the same pretrained vision–language backbone and a matched trainable-parameter budget for prompt tuning, adapter tuning, or lightweight residual tuning. The corresponding variant is denoted as OBSALIGN-VLM.

Table 5: **Dataset suite.** The primary metric follows the common protocol of each dataset family. "Client source" indicates how federated clients are constructed.

Dataset	Modalities	Task	Metric	Client source
UPMC Food-101 [Wang et al., 2015, Bossard et al., 2014]	Image, text	Food classification	Accuracy	Synthetic clients
MM-IMDb [Arevalo et al., 2017]	Poster image, plot text	Multi-label genre classification	Macro-F1	Synthetic clients
Hateful Memes [Kiela et al., 2020]	Image, text	Hate-speech detection	AUROC / Accuracy	Synthetic clients
CrisisMMD [Alam et al., 2018]	Image, text	Disaster-event classification	Macro-F1	Source/user grouping when available
MELD [Poria et al., 2019]	Audio, video, text	Emotion recognition	UAR / Macro-F1	Speaker/dialogue grouping
CREMA-D [Cao et al., 2014]	Audio, video	Emotion recognition	UAR	Speaker grouping
UCF101 [Soomro et al., 2012]	RGB video, audio / flow	Action recognition	Top-1 accuracy	Class-skewed synthetic clients
Moments-in-Time-10 [Monfort et al., 2019]	Video, audio	Action/event recognition	Top-1 accuracy	Class-skewed synthetic clients
EPIC-Kitchens [Damen et al., 2018]	RGB video, audio	Egocentric action recognition	Top-1 accuracy	Participant/video grouping
UCI-HAR [Anguita et al., 2013]	Accelerometer, gyroscope	Human activity recognition	Macro-F1 / Accuracy	Subject grouping
KU-HAR [Sikder and Nahid, 2021]	Accelerometer, gyroscope	Human activity recognition	Macro-F1 / Accuracy	Subject grouping
PTB-XL [Wagner et al., 2020]	ECG leads	ECG diagnosis	Accuracy / Macro-F1	Patient grouping
Sleep-EDF [Kemp et al., 2000]	PSG signals	Sleep-stage classification	Accuracy / Macro-F1	Subject grouping
Four-modal diagnostic subset (Fig. 4)	Image, Text, Audio, Sensor	8-category shared semantic classification	Prototype distance / semantic margin	Constructed multimodal sources

Dataset suite. Table 5 summarizes the benchmark suite. We include datasets that are commonly adopted in recent MMFL, missing-modality FL, and multimodal prompt-tuning studies. The suite covers image–text classification, social-media multimodal reasoning, audio–visual–text affective computing, action recognition, wearable-sensor activity recognition, and healthcare signal classification.

Data preprocessing. For image–text datasets, images are resized and normalized using the preprocessing pipeline of the chosen backbone, and text is tokenized by the corresponding text encoder. For foundation-model experiments, we use the tokenizer and image transform associated with the pretrained VLM. For action-recognition and affective-computing datasets, we follow the common feature-based protocol and use the same video, audio, and text features for all methods. For sensor and healthcare datasets, each sensor channel or signal lead is standardized using statistics computed on the training split only. Patient-, subject-, or speaker-level splits are disjoint across training, validation, and test sets. Validation data support model selection and hyperparameter tuning; test data are reserved for final evaluation.

Semantic units. ObsAlign requires semantic units $s \in \mathcal{S}$ for relation-level observability estimation. For single-label classification datasets, semantic units are the task labels. For multi-label datasets such as MM-IMDb, semantic units correspond to individual labels, and the assignment vector a_{ij} is the normalized multi-hot label vector. For datasets with soft or hierarchical labels, the semantic assignment function q_ω is fixed before federated training and is not updated during training. The same semantic units are used for ObsAlign, relation-conditioned evaluation, and the exploratory diagnostics.

Client construction. When natural client identities exist, we use them directly. Specifically, subjects define clients for wearable-sensor and sleep datasets, patients define clients for healthcare datasets, speakers or dialogues define clients for affective-computing datasets, and participants or video sources define clients for egocentric/action-recognition datasets when available. For datasets without natural client identities, we construct K clients using class-skewed Dirichlet partitioning:

$$\mathbf{p}_i \sim \mathrm{Dirichlet}(\alpha_{\mathrm{dir}}\mathbf{1}), \qquad \alpha_{\mathrm{dir}} \in \{0.1, 0.5, 1.0\}. \tag{48}$$

The default setting uses $\alpha_{\mathrm{dir}} = 0.5$. Smaller α_{dir} creates stronger label skew. Unless otherwise specified, image–text experiments use $K = 20$ clients. For datasets with natural clients, the number of clients is determined by the dataset. All baselines and ObsAlign use exactly the same client partitions under each random seed.

Incomplete-modality protocol. We assign each client a modality subset $\mathcal{M}_i \subseteq \mathcal{M}$ and represent actual input availability by $v_{ij,m}$, with $\mathcal{M}_{ij} = \{m \in \mathcal{M}_i : v_{ij,m} = 1\} \neq \varnothing$. Masks are generated once per seed and shared by all methods. Under client-level missingness, $v_{ij,m} = \mathbf{1}[m \in \mathcal{M}_i]$ for every sample. Under sample-level missingness, samples may miss additional modalities within $\mathcal{M}_i$ according to the same missing-rate budget. Fusion uses $\mathcal{M}_{ij}$; counts, prototypes, and dispersion use the corresponding observation mask. Pair-conditioned evaluation and co-observation diagnostics require $v_{ij,m}v_{ij,n} = 1$. We evaluate four test-time protocols:

- **Full**: all modalities are available at test time;

- **Same**: test-time modality availability follows the training mask distribution;
- **Single**: only one modality is available;
- **Severe**: at least half of the modalities are missing.

The **Same** protocol measures in-distribution incomplete-modality performance, while the **Severe** protocol measures robustness under stronger modality absence. Unless otherwise stated, the main comparison uses the **Same** protocol. The calibration study in Appendix C.2 uses the **Full**, **Single**, and **Severe** settings.

Backbone matching. All methods within the same regime use the same backbone. In the encoder-based regime, all methods use the same modality encoders or the same pre-extracted modality features. Generic FL baselines use the same multimodal prediction architecture as ObsAlign but remove the observability estimator, gated alignment loss, and bridge-aware aggregation. Feature-based MMFL baselines are implemented with their original loss terms and aggregation rules while keeping the backbone capacity fixed. In the foundation-model regime, all methods use the same pretrained VLM backbone, such as CLIP [Radford et al., 2021] or ViLT [Kim et al., 2021], depending on the dataset protocol. Text-only branches use BERT-style tokenization and embeddings [Devlin et al., 2019] when the compared method requires a separate text encoder.

ObsAlign variants. OBSALIGN-ENC is used in the encoder-based regime. It applies semantic observability estimation to modality-specific encoder features and uses observability-gated alignment between locally supported modality–semantic prototypes. OBSALIGN-VLM is used in the foundation-model regime. It keeps the pretrained VLM backbone matched with foundation-model baselines and applies ObsAlign to trainable prompts, adapters, or lightweight residual branches. In both variants, server-visible sketches are computed from the shared branch only, while private residual parameters remain local. The two variants share the same observability estimator, relation gate, semantic sketching mechanism, and bridge-aware aggregation rule.

Trainable-parameter matching. For foundation-model MMFL baselines, we match the trainable-parameter budget whenever possible. Prompt-based methods use the same prompt length, adapter-based methods use the same adapter rank or bottleneck width, and residual-branch methods use a comparable number of trainable parameters. If a baseline has an intrinsically different parameterization, we tune its trainable capacity to be within a small range of OBSALIGN-VLM. Figure 9 encodes the trainable-parameter count by bubble size.

Baseline implementation. We use official implementations whenever available. When official code is unavailable, we re-implement the method according to the paper and verify that the implementation reproduces the reported trend on the original benchmark whenever possible. For methods designed for a specific modality structure, we evaluate them only on compatible datasets. For example, HAR-specific methods are evaluated on sensor or activity-recognition datasets, while prompt-tuning methods are evaluated under the foundation-model regime. In the main comparison, all methods are trained under the same communication budget, local training budget, optimizer family, and modality mask unless the original method requires a different but necessary training component.

Comparison controls. We use the following controls throughout the experiments:

- **Same data split:** all methods use identical train/validation/test splits under the same seed.
- **Same client partition:** all methods use identical client assignments and client sample counts.
- **Same modality masks:** all methods observe the same missing-modality pattern during training and evaluation.
- **Same backbone:** methods within a comparison block use the same pretrained or non-pretrained backbone.
- **Same communication budget:** all methods use the same number of communication rounds and client participation rate.
- **Same local budget:** all methods use the same number of local epochs, batch size, and local optimizer steps.

Optimization. For feature-based or lightweight-encoder experiments, we use SGD with momentum 0.9, weight decay 10^{-5}, and a learning rate selected from $\{10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$. For transformer or foundation-model experiments, we use AdamW with learning rate selected from $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$. Unless otherwise specified, we use 200 communication rounds, one local epoch per round, batch size 32, and the same client participation rate for all methods. For larger foundation-model experiments, we use 250 communication rounds when this is required by the baseline protocol. The server learning rate is selected from $\{0.5, 1.0\}$ for FedAvg-style aggregation and is fixed across methods after validation selection.

ObsAlign hyperparameters. The default ObsAlign hyperparameters are

$$\lambda_{\text{align}} = 0.1, \quad \lambda_{\text{cal}} = 0.01, \quad \lambda_R = 10^{-4},$$

$$n_0 = 10, \quad \tau_v = 1.0, \quad \tau_l = 0.2, \quad \tau_h = 0.6, \quad \zeta = 0.5.$$

The observability coefficients in Eq. (6) are fixed across datasets as

$$\alpha = 1.0, \qquad \beta = 1.0, \qquad \eta = 0.5, \qquad \delta = 1.0, \qquad b_0 = 0.$$

We tune ObsAlign hyperparameters only on the validation split using the following grid:

$$\lambda_{\text{align}} \in \{0.01, 0.05, 0.1, 0.2\}, \quad \lambda_{\text{cal}} \in \{0, 0.01, 0.05\}, \quad \zeta \in \{0, 0.25, 0.5, 1.0\},$$

$$\tau_l \in \{0.1, 0.2, 0.3\}, \quad \tau_h \in \{0.5, 0.6, 0.7\}, \quad n_0 \in \{5, 10, 20\}, \quad \tau_v \in \{0.5, 1.0, 2.0\}.$$

We enforce $\tau_l < \tau_h$ in all runs. Once selected, the same hyperparameters are used across all test seeds for the corresponding dataset and comparison regime.

Warm start and semantic sketch refresh. ObsAlign requires an initial observability tensor O^0. Before the first communication round, clients compute initial semantic sketches using the initialized shared branch or the frozen pretrained backbone in the foundation-model regime. The server aggregates these sketches to obtain initial semantic prototypes and O^0. During training, round-t local optimization uses O^t, and sketches uploaded after round t are used only to compute O^{t+1}.

Relation-level evaluation. To evaluate whether ObsAlign improves weakly observed relations, we group modality–semantic relations (m, n, s) by their empirical bridge evidence $B_{m,n,s}$ defined in Eq. (41). Relations are divided into low-, medium-, and high-observability groups using tertiles of $B_{m,n,s}$ computed on the training split. We then report relation-conditioned performance:

$$\text{Score}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{(m,n,s) \in \mathcal{G}} \text{Score}_{m,n,s}, \qquad \mathcal{G} \in \{\mathcal{G}_{\text{low}}, \mathcal{G}_{\text{mid}}, \mathcal{G}_{\text{high}}\}. \tag{49}$$

These groups are used for post-hoc evaluation.

B.3 Computational Resources

All experiments were implemented in PyTorch and conducted on NVIDIA A100 GPUs. For encoder-based MMFL experiments, each training run used a single A100 GPU. For foundation-model experiments with ViLT-, CLIP-, or BERT-based backbones, each run used one to two A100 GPUs depending on the backbone size and dataset.

ObsAlign adds only lightweight server-side computation beyond standard federated training: client sketches scale as $\mathcal{O}(|\mathcal{M}_i|Sd)$, direct bridge estimation scales as $\mathcal{O}(|\mathcal{C}_t|M^2S)$, and multi-hop bridge estimation scales as $\mathcal{O}(SLM^3)$. Since the number of modalities is small in our benchmarks, the additional observability computation is minor compared with local model training.

C Additional Experimental Results

C.1 Additional Observability Diagnostics

Observability beyond marginal modality availability. We therefore ask whether relation-level observability explains alignment benefit better than marginal modality availability.

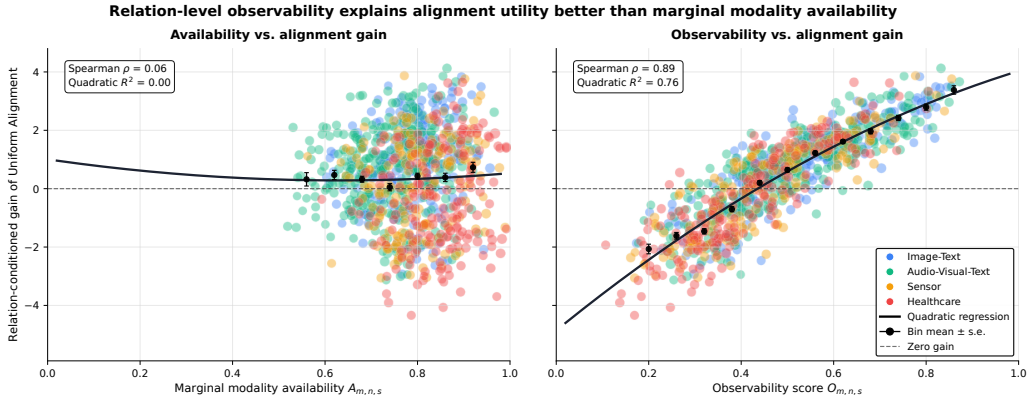

Figure 5: **Relation-level observability is more informative than marginal modality availability.** Each point denotes one modality–semantic relation $r = (m, n, s)$, averaged over three random seeds, and colored by dataset family. The vertical axis shows the relation-conditioned gain of Uniform Alignment over Task-only SourceAvg. **Left:** Marginal modality availability A_r has limited explanatory power: relations with similar availability still exhibit widely different alignment gains, including many negative-transfer cases. **Right:** Observability O_r shows a much clearer monotonic relation with alignment gain. The black curve shows a smoothed regression trend, and the black markers with error bars show bin-wise conditional means $\pm$ standard error.

For each modality–semantic relation $r = (m, n, s)$, we compare two predictors of relation-level alignment utility. The first is the *marginal modality availability*, using training-partition supports $n_{i,s} = \sum_{j \in \mathcal{D}_i^{\mathrm{tr}}} a_{ij,s}$ and $n_{i,m,s} = \sum_{j \in \mathcal{D}_i^{\mathrm{tr}}} a_{ij,s} v_{ij,m}$:

$$A_r = \frac{1}{2}\left(A_{m,s} + A_{n,s}\right), \qquad A_{m,s} = \frac{\sum_{i \in \mathcal{C}} n_{i,m,s}}{\sum_{i \in \mathcal{C}} n_{i,s} + \varepsilon}, \tag{50}$$

which measures how often each modality is individually visible for semantic unit s, regardless of whether the two modalities are co-observed. The second is the learned relation-level observability score $O_r = O_{m,n,s}$, which explicitly incorporates bridge evidence, consistency, support, and uncertainty.

As the target quantity, we use the relation-conditioned alignment gain of Uniform Alignment over Task-only SourceAvg:

$$G_r = \mathrm{Score}_r^{\mathrm{UA}} - \mathrm{Score}_r^{\mathrm{Task}}. \tag{51}$$

We then visualize the relation-level dependency using two side-by-side scatter plots. The left panel plots A_r against G_r, while the right panel plots O_r against G_r. Each point denotes one modality–semantic relation averaged over three random seeds, and colors indicate dataset families. To reveal the conditional trend, we overlay a smoothed regression curve together with bin-wise conditional means.

Figure 5 shows that marginal modality availability alone is a weak predictor of whether alignment is helpful. In the left panel, relations with similarly high availability still display large performance variation, and many of them even suffer negative transfer. In contrast, the right panel reveals a much cleaner monotonic relation between observability and alignment gain. In our diagnostic results, the Spearman correlation between marginal availability and gain is only 0.06, whereas the correlation between observability and gain reaches 0.89. A simple quadratic fit also explains substantially more variance with observability than with availability ($R^2 = 0.76$ vs. 0.00). Among relations with high marginal availability ($A_r > 0.70$), 34.6% still exhibit negative transfer, reflecting the importance of co-observation for reliable alignment. By contrast, among high-observability relations ($O_r > 0.70$), only 6.8% are harmful under uniform alignment. These results link alignment utility to the *bridge evidence* available for a specific modality–semantic relation.

Bridge-defined relation groups. The analyses above use the learned observability score to inspect the final relation structure. We also evaluate relation-level gains using groups defined by *training-time*

Table 6: Relation-conditioned gains under groups defined by training bridge evidence $B_{m,n,s}$. Groups are formed using only training-set bridge ratios, before evaluating final models. NTR denotes the negative-transfer rate, i.e., the fraction of relations whose gain over Task-only SourceAvg is negative. Values are averaged over datasets and three random seeds.

Group	Uniform Alignment			ObsAlign		
	Mean gain ↑	NTR ↓	10th pct. ↑	Mean gain ↑	NTR ↓	10th pct. ↑
Low bridge	-1.74 ± 0.18	89.6%	-3.42	$+0.28 \pm 0.09$	7.8%	-0.34
Mid bridge	$+0.42 \pm 0.11$	31.7%	-0.91	$+1.18 \pm 0.12$	9.6%	$+0.21$
High bridge	$+2.71 \pm 0.20$	5.9%	$+0.84$	$+3.76 \pm 0.21$	2.4%	$+1.62$

empirical bridge evidence $B_{m,n,s}$. Specifically, for each dataset and seed, we divide valid relations into three tertiles according to $B_{m,n,s}$ computed on the training split:

$$G^B_{\text{low}}, \quad G^B_{\text{mid}}, \quad G^B_{\text{high}}.$$

These bridge-defined groups are fixed before training the evaluated models. We then report the relation-conditioned gain over Task-only SourceAvg and the negative-transfer rate within each bridge-defined group.

Table 6 shows that Uniform Alignment harms weakly bridged relations under training-evidence grouping, with an average gain of -1.74 and an 89.6% negative-transfer rate. ObsAlign substantially reduces this failure mode, increasing the low-bridge mean gain to $+0.28$ and lowering the negative-transfer rate to 7.8%. In the high-bridge regime, ObsAlign preserves and further improves the positive effect of alignment. The gate uses relation evidence to suppress alignment where training sources provide weak bridge support.

Importance of observability components. ObsAlign defines relation-level observability using four ingredients: bridge evidence, cross-modal consistency, side support, and prototype uncertainty. We use a post-hoc contribution analysis to quantify each component's relative importance and direction of association with alignment gain.

For each modality–semantic relation $r = (m, n, s)$, we extract the four relation-level observability components computed at the final communication round:

$$x_r = \big[\text{Bridge}_r,\ \text{Cons}_r,\ \text{Supp}_r,\ \text{Unc}_r\big]. \tag{52}$$

As the prediction target, we use the relation-conditioned alignment gain of Uniform Alignment over Task-only SourceAvg:

$$G_r = \text{Score}_r^{\text{UA}} - \text{Score}_r^{\text{Task}}. \tag{53}$$

We analyze the contribution of the four components using two complementary post-hoc models.

First, we fit a nonlinear gradient-boosted regressor $f(\cdot)$ to predict G_r from x_r, and compute SHAP values to quantify feature contribution at the relation level. The SHAP beeswarm plot therefore reveals both the overall importance of each component and the direction in which large or small feature values affect the gain.

Second, we fit a standardized linear surrogate model

$$\tilde{G}_r = \beta_0 + \beta_{\text{B}}\,\widetilde{\text{Bridge}}_r + \beta_{\text{C}}\,\widetilde{\text{Cons}}_r + \beta_{\text{S}}\,\widetilde{\text{Supp}}_r + \beta_{\text{U}}\,\widetilde{\text{Unc}}_r + \epsilon_r, \tag{54}$$

where all predictors are z-score standardized. We report bootstrapped 95% confidence intervals for the standardized coefficients. Together, the SHAP beeswarm plot and the coefficient forest plot reveal both the *importance ranking* and the *signed direction* of each observability component.

Figure 6 shows a clear importance hierarchy among the four observability components. The SHAP analysis ranks the components as

$$\text{Bridge} > \text{Unc} > \text{Cons} > \text{Supp},$$

with mean absolute SHAP values of 0.83, 0.42, 0.33, and 0.17, respectively. Larger bridge evidence is associated with higher alignment gain, while larger uncertainty is associated with lower gain.

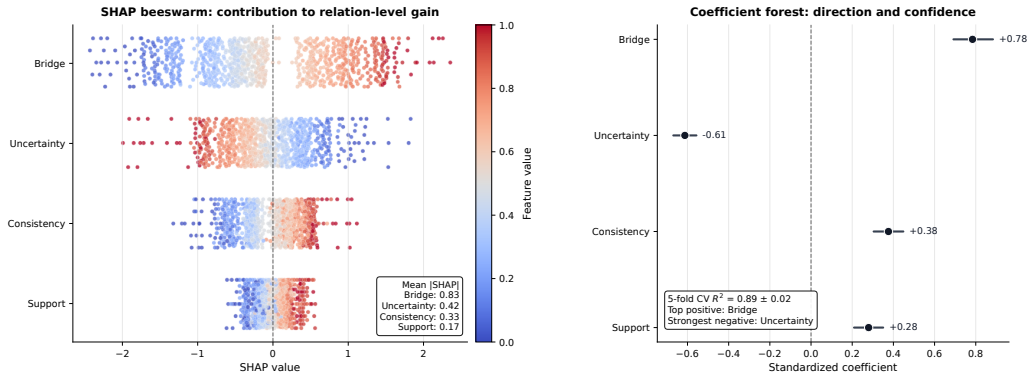

Figure 6: **Contribution analysis of the four observability components.** The target is the relation-conditioned gain of Uniform Alignment over Task-only SourceAvg. **Left:** SHAP beeswarm plot for a nonlinear regressor fitted on relation-level gain. Each point denotes one modality–semantic relation. The horizontal axis is the SHAP contribution, the color encodes the feature value, and the features are sorted by mean absolute SHAP value. **Right:** Coefficient forest plot for a standardized linear surrogate with bootstrapped 95% confidence intervals. Bridge evidence is the strongest positive predictor, uncertainty is the strongest negative predictor, consistency provides an additional positive signal, and support has a weaker but still positive effect.

Consistency provides a smaller positive contribution by refining plausible relations. Side support has a weaker positive effect and becomes useful when accompanied by bridge evidence.

The coefficient forest plot yields the same conclusion. The standardized coefficients are $\beta_B = +0.78$, $\beta_U = -0.61$, $\beta_C = +0.38$, and $\beta_S = +0.28$, each with a 95% confidence interval excluding zero. This means that bridge evidence is the dominant positive driver of alignment utility, while uncertainty is the dominant negative driver. The nonlinear regressor achieves a 5-fold cross-validated R^2 of 0.89 ± 0.02, indicating that the four components explain a substantial portion of the variation in relation-level gain. Bridge evidence and uncertainty provide the strongest signals, complemented by consistency and side support.

Negative transfer mitigation for weakly observable relations. We next test whether ObsAlign directly protects weakly observable relations from harmful alignment. We divide all modality–semantic relations into low-, medium-, and high-observability groups and compare their gain distributions under Uniform Alignment and ObsAlign in Figure 7. Uniform Alignment produces a long negative tail in the low-observability group, with a mean gain of -2.53 and a negative-transfer rate of 97.7%. ObsAlign shifts this distribution upward, increasing the mean gain to $+0.21$ and reducing the negative-transfer rate to 15.0%, while preserving stronger gains in the high-observability regime. The gate suppresses unreliable alignment while retaining gains on well-observed relations, explaining the large drop when it is removed in Table 3.

C.2 Missing-Modality Calibration

We evaluate the contribution of calibration to confidence under missing or weakly observable modalities by comparing ObsAlign with a variant without the calibration loss, denoted w/o calibration. We evaluate both models under three inference conditions: Full-modal, Single-modal, and Severe-missing.

We use expected calibration error (ECE) as the calibration metric. Let $c_j = \max_k \hat{p}_{j,k}$ be the predicted confidence of sample j, and let $\{B_b\}_{b=1}^{M}$ be confidence bins. ECE is defined as

$$\mathrm{ECE} = \sum_{b=1}^{M} \frac{|B_b|}{N} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right| .$$

Lower ECE indicates better calibration.

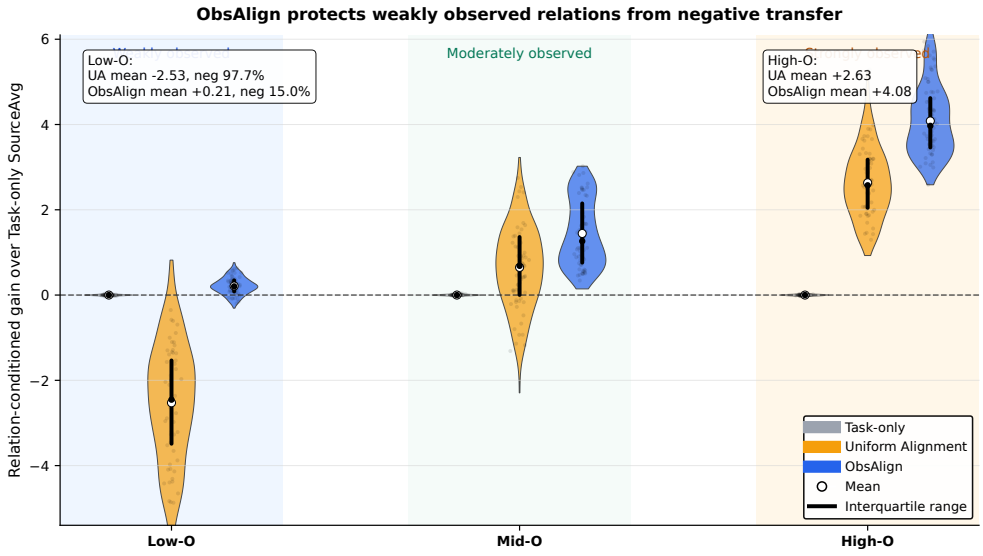

Figure 7: **Negative transfer mitigation for weakly observable relations.** Relations are grouped by observability level, and the vertical axis shows gain over Task-only SourceAvg. ObsAlign sharply reduces the long negative tail caused by Uniform Alignment in the low-observability group while preserving high-observability gains.

Table 7: Missing-modality calibration. Accuracy and ECE are averaged over benchmark groups and three random seeds. The calibration loss has a small effect under full modalities but becomes substantially more important as modality evidence becomes incomplete.

Evaluation	w/o calibration		ObsAlign		ECE reduction
	Acc. ↑	ECE ↓	Acc. ↑	ECE ↓	
Full-modal	84.7 ± 0.3	3.8 ± 0.4	85.2 ± 0.3	2.4 ± 0.3	36.8%
Single-modal	73.1 ± 0.5	9.7 ± 0.8	74.5 ± 0.4	4.9 ± 0.5	49.5%
Severe-missing	61.8 ± 0.6	16.7 ± 1.1	64.1 ± 0.5	6.2 ± 0.6	62.9%

Table 7 shows that calibration is most important under modality absence. Under Full-modal evaluation, both models are reasonably calibrated, and ObsAlign slightly improves both accuracy and ECE. Under Single-modal and Severe-missing evaluation, the model without calibration becomes increasingly overconfident, leading to much higher ECE. ObsAlign reduces ECE by 49.5% under Single-modal evaluation and by 62.9% under Severe-missing evaluation, while also improving accuracy. The calibration term improves accuracy and reduces spuriously confident predictions on weakly observable missing-modality relations.

C.3 Hyperparameter Sensitivity and Communication Efficiency

Hyperparameter sensitivity. We further analyze the sensitivity of ObsAlign to five important hyperparameters, including the alignment weight λ_{align}, the calibration weight λ_{cal}, the bridge-aware aggregation strength ζ, the observability gate interval (τ_l, τ_h), and the reliability temperature τ_v. For each study, we vary one hyperparameter while keeping all others at their default values. The default setting follows the main experiments: $\lambda_{\text{align}} = 0.1$, $\lambda_{\text{cal}} = 0.01$, $\zeta = 0.5$, $(\tau_l, \tau_h) = (0.2, 0.6)$, and $\tau_v = 1.0$. Figure 8 summarizes the mean and standard deviation over three random seeds.

As shown in Figure 8, ObsAlign is stable across a broad range of hyperparameter choices. The alignment weight λ_{align} exhibits a mild unimodal trend: too small a value underuses reliable cross-modal evidence, while too large a value can over-align weakly supported relations. The calibration weight λ_{cal} mainly affects severe-missing robustness, indicating that calibration is especially helpful when modality evidence is incomplete. The bridge coefficient ζ is also important: $\zeta = 0$ degenerates to

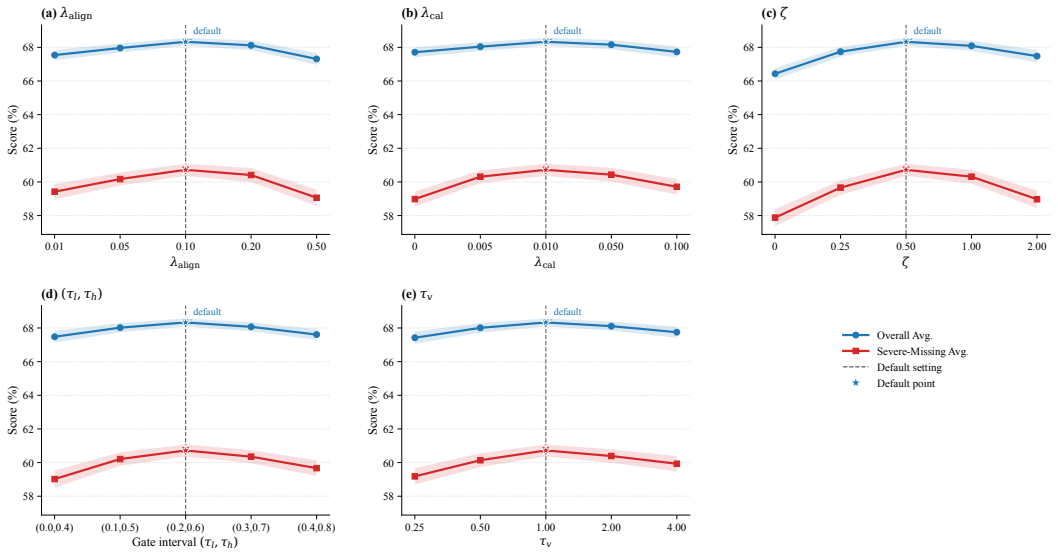

Figure 8: **Parameter sensitivity analysis of ObsAlign.** We vary one hyperparameter at a time while keeping the others fixed at their default values. Each subplot reports the overall average score and the severe-missing average score, with shaded bands indicating mean $\pm$ standard deviation over three random seeds. The dashed vertical line marks the default setting in each panel. ObsAlign remains stable under moderate changes in alignment strength, calibration strength, bridge-aware aggregation, gate thresholding, and sketch reliability temperature.

sample-count aggregation and leads to a clear performance drop, whereas overly large values slightly hurt performance by overemphasizing bridge clients. The gate interval (τ_l, τ_h) remains robust around the default range, but overly permissive or overly conservative thresholds both degrade performance. Finally, τ_v admits a range of effective reliability temperatures, provided sketch uncertainty is neither over-penalized nor ignored. Overall, these results indicate that the default configuration provides a good balance between alignment strength, observability selectivity, and robustness to missing modalities.

Communication–performance tradeoff. Beyond predictive performance, multimodal federated learning must also be communication-efficient. This is particularly important in incomplete-modality MMFL, where some recent methods rely on prompt tuning, teacher guidance, or large foundation-model updates, which can substantially increase the amount of uploaded information in each communication round. We therefore ask whether ObsAlign offers a better communication–performance tradeoff than existing MMFL baselines.

We measure the total client-to-server uploaded communication over the full training process. For method a, we define

$$\mathcal{C}_{\text{up}}^{(a)} = \sum_{t=1}^{T} \sum_{i \in \mathcal{C}_t} \left(\text{bytes}\left(\Delta\Theta_{i,t}^{(a)} \right) + \text{bytes}\left(\mathcal{U}_{i,t}^{(a)} \right) \right), \tag{55}$$

where $\Delta\Theta_{i,t}^{(a)}$ is the uploaded model update and $\mathcal{U}_{i,t}^{(a)}$ denotes any additional uploaded information, such as prototypes, prompts, teacher statistics, missing-modality tokens, or semantic sketches. For ObsAlign, $\mathcal{U}_{i,t}$ consists only of the compact semantic sketch introduced in Eqs. (3)–(4). We report $\mathcal{C}_{\text{up}}^{(a)}$ in gigabytes (GB) and plot it against the average test score across benchmark groups. Following the main results section, the average score is computed within each compatible comparison regime. To make the tradeoff visually interpretable, we use a Pareto bubble plot: the x-axis is total uploaded communication, the y-axis is average score, and the bubble size represents the number of trainable parameters. Methods on the upper-left frontier are preferable, as they achieve higher performance with lower communication.

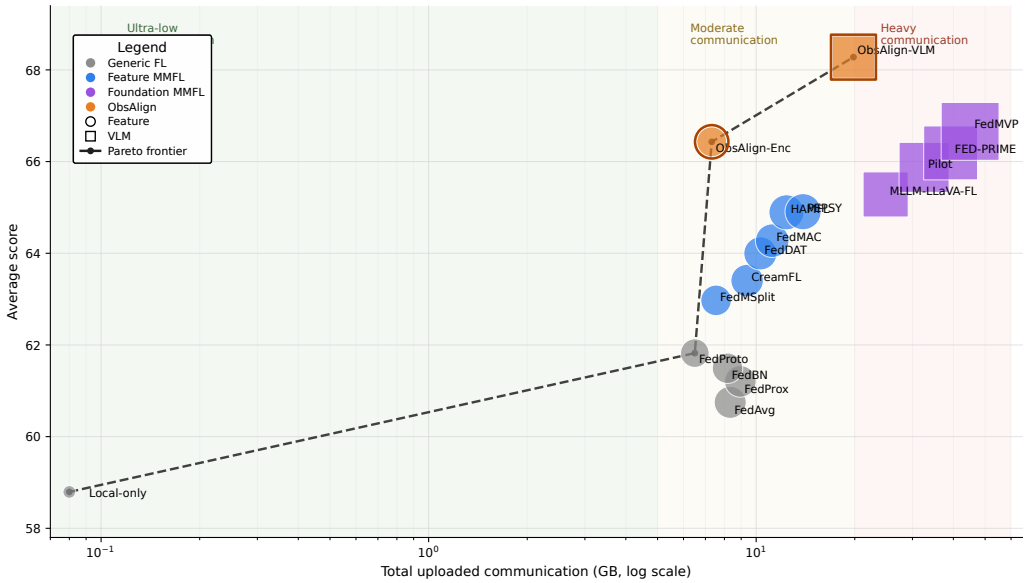

Figure 9: **Communication–performance Pareto comparison across MMFL methods.** The x-axis shows total uploaded communication over the full training run (log scale), and the y-axis shows the average score in the corresponding comparison regime. Bubble size indicates the number of trainable parameters. Colors denote method families. The dashed line connects Pareto-efficient methods. ObsAlign lies near the upper-left frontier: it achieves stronger or comparable performance with substantially lower uploaded communication than heavier prompt-, teacher-, and foundation-model-based baselines. This suggests that uploading compact semantic sketches is substantially more communication-efficient than transmitting larger adaptation states or auxiliary teacher information.

Figure 9 shows that ObsAlign achieves a favorable communication–performance tradeoff. Among feature-based methods, OBSALIGN-ENC reaches the best average score while requiring only moderate uploaded communication, with lower communication cost than several prototype- and auxiliary-transfer baselines. Among foundation-model-compatible methods, OBSALIGN-VLM achieves the highest average score while using substantially less uploaded communication than heavy prompt- or teacher-style approaches such as Pilot, FED-PRIME, and FedMVP. ObsAlign lies close to the Pareto frontier. Compact semantic sketches support relation-level coordination with low communication overhead.

C.4 Pre-Alignment Intervention Validation

We compute every ordering signal from the same Task-only warm-start checkpoint before any gated update. Each relation $r = (m, n, s)$ is then aligned alone for the same short update budget. Its intervention gain is

$$\Delta_r^{\text{int}} = \text{Score}(\text{align only } r) - \text{Score}(\text{Task-only}). \tag{56}$$

Spearman correlation measures whether a signal orders these intervention gains, and AUROC measures discrimination of $\Delta_r^{\text{int}} > 0$. Table 8 reports the three diagnostic datasets. The mean Spearman/AUROC is $0.75/0.91$ for O^0, $0.57/0.81$ for Bridge-only, $0.25/0.66$ for Side support, and $0.07/0.54$ for Marginal availability.

C.5 Targeted Selection under Matched Alignment Budgets

All selectors use the same splits, backbones, masks, optimization budgets, and evaluation protocols. Uniform-Budget scales a global alignment coefficient to match ObsAlign's total alignment mass and keeps 100% coverage. Bridge-TopK and ObsAlign-Hard additionally match ObsAlign's active-relation coverage of 54.2%. Table 4 reports Overall and Severe using the five encoder-based benchmark-family averages in Table 3. This aggregate differs from the 13-dataset mean that combines matched VLM and encoder variants in Table 16.

Table 8: Pre-intervention ordering versus single-relation alignment gain. Top-30% is the mean intervention gain among the highest-ranked 30% of relations.

Dataset	Ordering signal	$\rho \uparrow$	AUROC $\uparrow$	Top-30% gain $\uparrow$
MM-IMDb	Random	0.01	0.50	+0.08
MM-IMDb	Marginal availability	0.07	0.54	+0.22
MM-IMDb	Side support	0.24	0.65	+0.68
MM-IMDb	Bridge-only	0.52	0.78	+1.46
MM-IMDb	ObsAlign O^0	0.71	0.88	+2.18
MELD	Random	-0.02	0.49	+0.03
MELD	Marginal availability	0.10	0.56	+0.31
MELD	Side support	0.29	0.68	+0.84
MELD	Bridge-only	0.58	0.81	+1.72
MELD	ObsAlign O^0	0.76	0.91	+2.54
PTB-XL	Random	0.00	0.50	+0.05
PTB-XL	Marginal availability	0.05	0.53	+0.19
PTB-XL	Side support	0.22	0.64	+0.74
PTB-XL	Bridge-only	0.61	0.83	+1.96
PTB-XL	ObsAlign O^0	0.79	0.93	+3.08

Table 9: Selective feature alignment on the compatible subset only.

Method	Subset avg. $\uparrow$	Severe $\uparrow$	Low-O gain $\uparrow$	NTR $\downarrow$
Uniform-Budget	64.88	55.31	-0.82	70.8%
FedSEA	65.84	56.82	+0.04	24.1%
Bridge-TopK	66.31	58.04	+0.15	14.6%
ObsAlign	67.58	59.33	+0.31	4.2%

Low-O NTR is the fraction of diagnostic relations whose gain over Task-only is negative. Bridge-TopK uses direct and multi-hop bridge evidence.

FedSEA [Tan et al., 2023] is evaluated on its structurally compatible dataset subset. Table 9 reports averages within this subset. ObsAlign improves the subset average by 1.74 points over FedSEA.

C.6 Misleading Bridge Evidence

We hold marginal support and co-observation counts fixed while corrupting bridge evidence. Semantic mismatch pairs modality n with an incorrect semantic unit $s' \neq s$ within selected bridge sources. Domain shortcuts introduce a stable relation in a source/domain subgroup that is unsupported by other sources. A false-high relation is a corrupted relation with $O_r > \tau_h$.

Table 10 reports the 20% corruption study. Consistency is especially useful for semantic mismatch, while cross-source Uncertainty is especially useful for domain shortcuts. ObsAlign retains clean high-O gains of $+3.88/+3.89$, while corrupted-relation gains remain slightly negative at $-0.05/-0.08$. Table 11 reports the more severe 40% setting.

Shared spurious relations. Detecting a spurious relation shared by all sources requires counter-domain evidence, interventions, or additional supervision. The evaluated corruption settings expose inconsistencies within or across sources.

C.7 Score Sensitivity and Default Settings

Table 12 complements Figure 8 by varying the score coefficients one at a time. The fixed cross-dataset setting is $(\alpha, \beta, \eta, \delta, b_0) = (1, 1, 0.5, 1, 0)$. Overall and Severe follow Table 3. The tested settings within 0.5 Overall points of the default include $\alpha, \beta, \delta \in [0.75, 1.50]$, $\eta \in [0.25, 0.75]$, and $b_0 \in [-0.25, 0.25]$.

The default recipe fixes these coefficients, computes O^0 through the warm-start sketch pass, and uses validation data for a small operating-point grid: $\lambda_{\text{align}} \in \{0.05, 0.1, 0.2\}$, $\tau_l \in \{0.1, 0.2, 0.3\}$, $\tau_h \in \{0.5, 0.6, 0.7\}$, and $\zeta \in \{0.25, 0.5, 1.0\}$. This is a compact choice within the broader validation grid in Appendix B.2. The score coefficients remain fixed during this operating-point search.

Table 10: Misleading bridges at 20% corruption. False-high is the fraction of corrupted relations with $O_r > \tau_h$. NTR and gain are evaluated on corrupted relations.

Corruption	Method	False-high ↓	Gain ↑	NTR ↓	Overall ↑
Mismatch	Bridge-only	52.4%	−1.28	72.4%	64.82
Mismatch	w/o Consistency	41.2%	−0.88	62.7%	65.48
Mismatch	w/o Uncertainty	35.9%	−0.66	52.8%	65.91
Mismatch	ObsAlign	14.7%	−0.05	22.1%	67.25
Shortcut	Bridge-only	57.4%	−1.31	73.5%	64.56
Shortcut	w/o Consistency	34.8%	−0.66	52.9%	65.61
Shortcut	w/o Uncertainty	46.2%	−0.92	65.1%	65.18
Shortcut	ObsAlign	11.6%	−0.08	20.7%	67.12

Table 11: Stress test at 40% bridge corruption.

Corruption	Method	False-high ↓	Gain ↑	NTR ↓	Overall ↑
Mismatch	Bridge-only	73.1%	−2.44	88.6%	63.10
Mismatch	ObsAlign	25.8%	−0.42	39.7%	66.08
Shortcut	Bridge-only	76.8%	−2.26	88.9%	62.97
Shortcut	ObsAlign	23.4%	−0.41	39.8%	65.91

Table 13 reports the statistics entering the sigmoid, before multiplication by the bridge-sufficiency factor. These are the raw formula values and their coefficient-weighted values, without an additional normalization step. Bridge is the strongest positive predictor in the component analysis, and Uncertainty is the strongest negative predictor.

Table 12: One-factor coefficient sensitivity. Values marked default use the fixed cross-dataset score setting.

Coefficient	Value	Overall ↑	Severe ↑	Low-O NTR ↓
α (Bridge)	0.50	67.61	59.66	2.8%
α (Bridge)	0.75	68.08	60.27	3.0%
α (Bridge)	1.00 (default)	68.34	60.72	3.4%
α (Bridge)	1.50	68.11	60.36	5.9%
α (Bridge)	2.00	67.55	59.61	10.8%
β (Consistency)	0.50	67.72	59.88	6.6%
β (Consistency)	0.75	68.10	60.35	4.6%
β (Consistency)	1.00 (default)	68.34	60.72	3.4%
β (Consistency)	1.50	68.18	60.58	3.1%
β (Consistency)	2.00	67.86	60.07	2.9%
η (Support)	0.25	67.89	60.08	3.2%
η (Support)	0.50 (default)	68.34	60.72	3.4%
η (Support)	0.75	68.16	60.50	4.0%
η (Support)	1.00	67.88	60.02	5.7%
η (Support)	1.50	67.29	59.21	9.4%
δ (Uncertainty)	0.50	67.52	59.44	10.5%
δ (Uncertainty)	0.75	68.05	60.20	6.0%
δ (Uncertainty)	1.00 (default)	68.34	60.72	3.4%
δ (Uncertainty)	1.50	68.12	60.51	2.6%
δ (Uncertainty)	2.00	67.64	59.78	1.9%
b_0	−0.50	67.28	59.10	13.2%
b_0	−0.25	67.93	60.02	7.2%
b_0	0.00 (default)	68.34	60.72	3.4%
b_0	0.25	68.11	60.44	2.5%
b_0	0.50	67.57	59.63	1.7%

Table 13: Raw and coefficient-weighted component scales before the sigmoid. Brackets give the interquartile range.

Term	Raw median [IQR]	Coefficient	Weighted median [IQR]
Bridge	0.93 [0.38, 1.46]	1.0	0.93 [0.38, 1.46]
Consistency	0.68 [0.31, 0.89]	1.0	0.68 [0.31, 0.89]
Support	2.42 [1.48, 3.36]	0.5	1.21 [0.74, 1.68]
Uncertainty	0.84 [0.57, 1.15]	1.0	0.84 [0.57, 1.15]

C.8 Sketch Communication, Runtime, and Empirical Leakage

Each visible modality–semantic pair uploads a d-dimensional prototype and three scalar statistics (count, dispersion, and reliability). For float32 arrays, the exact payload is

$$C_i^{\text{sketch}} = 4|\mathcal{M}_i|S(d+3) \text{ bytes per client-round.} \tag{57}$$

Table 14 uses KiB ($1\ \text{KiB} = 1024$ bytes) and averages the visible-modality count over clients in each dataset configuration. Sketch overhead is the incremental sketch upload divided by the corresponding model-update upload, $100C_i^{\text{sketch}}/C_i^{\text{model}}$. It excludes protocol headers and download traffic. The mean of the displayed per-dataset overheads is 0.24%, the maximum is 0.98%, and server refresh averages 4.7 ms per round (maximum 18.6 ms). The arithmetic mean of the displayed sketch payloads is 28.33 KiB per client-round.

Counts and prototypes can expose source-level semantic proportions and domain statistics. Table 15 evaluates a distribution-inference attack against uploaded sketches. Lower R^2 and higher MAE indicate weaker empirical inference. Noisy prototypes are L2-clipped and perturbed with Gaussian noise of standard deviation 0.05. Log-binned counts and clipped/noisy prototypes reduce attack R^2 from 0.71 to 0.23, with an Overall decrease from 68.34 to 67.58. Formal differential privacy requires a specified adjacency relation, sensitivity analysis, and privacy accounting.

Table 14: Float32 array payloads and server refresh time. Upload sizes are KiB per client-round. The overhead denominator is the model update only.

Dataset	Sketch KiB	Model KiB	Overhead	Refresh ms
UPMC	158.38	30,720	0.52%	18.6
MM-IMDb	36.07	30,720	0.12%	5.2
Hateful Memes	3.14	30,720	0.01%	1.0
CrisisMMD	12.95	30,720	0.04%	1.7
MELD	7.52	8,192	0.09%	2.4
CREMA-D	4.76	8,192	0.06%	1.2
UCF101	80.11	8,192	0.98%	9.3
MiT10	7.93	8,192	0.10%	1.5
EPIC	16.38	8,192	0.20%	2.6
UCI-HAR	4.76	3,072	0.15%	0.8
KU-HAR	14.28	3,072	0.46%	1.4
PTB-XL	16.63	6,144	0.27%	14.2
Sleep-EDF	5.37	6,144	0.09%	1.4

Table 15: Empirical sketch leakage–utility tradeoff.

Sketch variant	Attack $R^2 \downarrow$	MAE $\uparrow$	Overall $\uparrow$	Severe $\uparrow$
Exact sketch	0.71	0.071	68.34	60.72
Coarsened count	0.43	0.109	68.18	60.45
Binary support	0.21	0.161	67.61	59.63
Noisy prototype	0.47	0.101	67.95	60.11
Coarsened + noisy	0.23	0.154	67.58	59.70

C.9 Natural and Constructed Sources

Source identity and modality missingness are separate axes. The natural/mixed group comprises CrisisMMD, MELD, CREMA-D, EPIC-Kitchens, UCI-HAR, KU-HAR, PTB-XL, and Sleep-EDF. These use user/source, speaker/dialogue, participant/video, subject, or patient grouping. CrisisMMD is labeled mixed because source/user grouping is used where available. The constructed group comprises UPMC Food-101, MM-IMDb, Hateful Memes, UCF101, and MiT10 and uses Dirichlet/class-skewed partitions.

Both groups use controlled modality masks, with available-modality structure retained where applicable. Every method receives the same source assignments, masks, and seeds. Table 16 averages each dataset against its strongest matched baseline, using ObsAlign-VLM for the four image–text datasets and ObsAlign-Enc for the other nine. These summaries average dataset-specific metrics across benchmarks.

Table 16: Results grouped by source identity. Gains are points in the dataset-specific metric. Means and gains are rounded independently. Controlled missingness applies to both groups.

Source identity	# datasets	Baseline	ObsAlign	Gain ↑
Natural/mixed	8	67.25	69.43	+2.19
Constructed	5	62.77	64.70	+1.93
All datasets	13	65.53	67.61	+2.09

D Broader Impacts

ObsAlign may improve the reliability of multimodal federated learning by reducing unsupported cross-modal alignment, negative transfer, and over-aggregation from clients that do not provide useful bridge evidence. Potential positive impacts include more reliable learning in healthcare, mobile sensing, affective computing, and vision–language applications where sources have incomplete or heterogeneous modality access. Potential harms include disclosure of source-level semantic distributions through sketches, unequal performance for underrepresented semantic groups, and overconfident transfer under misleading evidence. Source-level auditing, conservative gating, access controls, and application-specific validation remain necessary before deployment.

E LLM Usage

LLM-based tools assisted with wording, grammar, LaTeX formatting, code snippets, debugging, and note organization. The authors specified and verified the scientific content and reviewed all assisted text and code before use.

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 abstract and introduction describe relation-level, evidence-dependent alignment in incomplete-modality MMFL. Section 4 defines semantic sketches, observability estimation, gated alignment, and bridge-aware aggregation; Section 5 evaluates these components.

   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: The Limitations paragraph discusses semantic-unit dependence, shared spurious relations, controlled missingness, and privacy requirements. Appendices C.6, C.8, and D give the corresponding failure and impact analyses.

   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: [N/A].

Justification: The contribution is algorithmic and empirical. The mathematical sections define the observability estimator, gated alignment objective, bridge-aware aggregation rule, and computational complexity.

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: The paper describes the benchmark suite, client construction, modality-missing protocols, evaluation metrics, baseline groups, backbone matching, communication rounds, local epochs, optimizers, hyperparameter grids, random seeds, and relation-level evaluation protocols. The method section and Algorithm specify the main training procedure needed to reproduce ObsAlign.

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: Code and usage instructions are available at `https://github.com/Xiaocanran1230/ObsAlign`. Benchmark-specific data preparation and reproduction scripts are not included. Appendix B.2 documents implementation, optimization, and evaluation protocols for the existing benchmarks.

   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: The paper specifies dataset groups, evaluation metrics, client construction, incomplete-modality protocols, backbone matching, trainable-parameter matching, fair-comparison controls, optimizers, learning-rate grids, communication rounds, local epochs, batch size, ObsAlign hyperparameters, and validation-based hyperparameter selection in the main text and appendix.

   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: The paper reports results averaged over three random seeds and states the random-seed protocol in the experimental setup. Additional analyses report uncertainty information where appropriate, including mean $\pm$ standard deviation in the hyperparameter sensitivity study and bootstrapped 95% confidence intervals in the observability component analysis.

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: Appendix B.3 describes the compute setting.

   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 work uses existing benchmarks and standard experimental protocols for multimodal federated learning. It involves no new human-subject data collection or release of high-risk models or scraped datasets.

   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: Appendix D discusses more reliable learning, possible leakage of semantic distributions, unequal performance for underrepresented groups, and overconfident transfer under misleading evidence, together with potential mitigations.

    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 does not release a new pretrained model, image generator, scraped dataset, or other high-risk asset requiring special release safeguards. The experiments use existing benchmarks and the proposed method is an algorithmic framework for multimodal federated learning.

    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: [No]

    Justification: The paper credits the existing datasets, baselines, and pretrained backbones through citations to their original publications.

    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 introduce or release a new dataset, benchmark, or other standalone asset. The contribution is a new algorithmic framework and evaluation protocol built on existing benchmarks.

    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 paper does not involve crowdsourcing experiments or new research with 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 paper does not involve new human-subject research, crowdsourcing, or collection of participant data, so IRB approval is not applicable.

    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: Appendix E describes assistive use of LLM-based tools for text, formatting, and coding. The authors specified and verified the scientific content and reviewed all assisted text and code.

    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.