LEARNING CROSS-MODAL GENERATIVE PRIORS FOR MULTIMODAL FEDERATED LEARNING WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal federated learning often involves clients that observe only one modality, while genuine image–text supervision is concentrated at paired clients. We present FedCGP, a framework that learns conditional feature priors from paired observations and uses them to regularize incomplete clients. Its central idea is to train the representation map actually used during local learning: fitting a denoiser on corrupted targets alone does not ensure useful outputs when queries are instead seeded by independent Gaussian noise. Paired clients therefore combine real-pair alignment, conditional denoising, and contrastive guidance through a single-step query. Incomplete clients keep the prior fixed, detach its outputs, and update their observed encoders alongside real supervised objectives. Block-specific aggregation preserves paired-only prior learning while balancing paired and incomplete contributions to shared representations. We analyze how guidance changes normalized features and how auxiliary gradients interact with real-task learning. Experiments on Flickr30k and MS-COCO combine federation-level evaluation with controlled generator and query-path studies. In the frozen-feature study, generated-to-real supervision substantially improves denoiser correspondence, while deterministic translation remains a strong alternative. Replacing the trained single-step query with a different query path at fixed weights causes a large drop in correspondence, showing that the query procedure is part of the learned interface. These results highlight query-aligned learning and selective update ownership as key design choices for transferring paired supervision under missing modalities.

1 INTRODUCTION

Federated learning trains models across clients while keeping observations local (McMahan et al., 2017; Kairouz et al., 2021). In multimodal applications, clients also differ in the modalities they observe: some possess image–text pairs, whereas others have only images or text (Che et al., 2023; Feng et al., 2023). This asymmetry affects the supervision available for learning. Paired clients can associate visual and textual content; an image-only client may have useful class labels but no matching description. A multimodal federation should draw on both sources of information, allowing incomplete clients to benefit from paired supervision while continuing to learn from their own observations.

Existing methods connect clients through representation transfer, complementary model collaboration, or missing-feature prediction (Yu et al., 2023; Sun et al., 2024; Xiong et al., 2023; Poudel et al., 2025). Sharing model parameters allows information to cross client boundaries, but does not by itself supply the complementary feature needed for local cross-modal learning. A conditional prior offers a direct route: learn modality relations from genuine pairs, then query those relations using the modality available at an incomplete client. The generated feature becomes a local auxiliary target. Operating in representation space also lets clients use this target without reconstructing a missing image or sentence.

Making this transfer useful requires connecting how the prior is trained to how it is queried. Conditional denoising learns from a corrupted real target together with its observed counterpart. An incomplete client has no complementary target to corrupt, so its query must start from independent

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

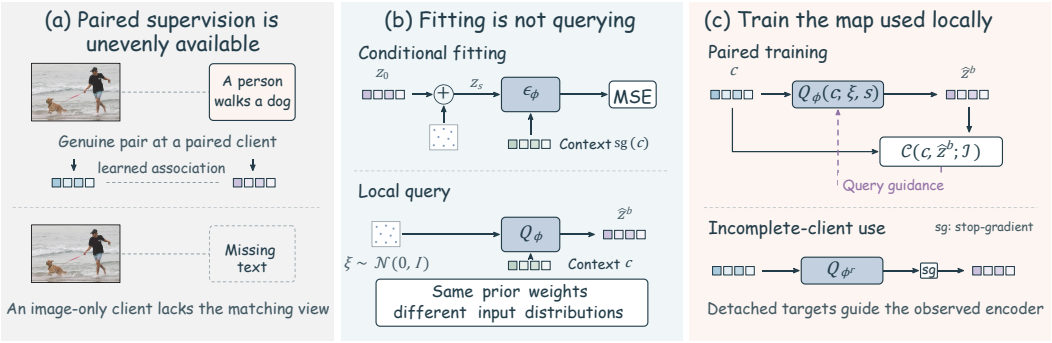


Figure 1: Learning and transferring paired supervision. The missing-view example illustrates client-level modality incompleteness.

Gaussian noise. At an interior noise level, these inputs follow different distributions. Moreover, a denoising objective measures noise prediction, whereas local cross-modal learning depends on the direction and correspondence of the resulting feature. A prior can therefore fit paired data well without producing a useful target through the query used locally.

A second issue concerns who should update the prior. Paired clients observe both sides of a cross-modal relation and can supervise conditional learning directly. Incomplete clients observe only one side; a generated target carries information from the learned prior but supplies no new paired observation. These clients nevertheless provide valuable gradients for their observed encoders through real task labels. The federation must preserve that contribution while keeping conditional learning tied to the clients that possess paired evidence. This requires separate aggregation rules for representations and the prior.

We introduce FedCGP, a framework that learns cross-modal generative priors with explicit query guidance and update ownership (Figure 1). Paired clients combine real-pair alignment and conditional denoising with contrastive guidance through the same Gaussian-seeded, single-step map used by incomplete clients. The latter keep the received prior fixed and detach its outputs, using the generated feature to regularize their observed encoder alongside real supervision. The server averages prior updates from paired contributors and balances paired and incomplete updates to compatible representation blocks. This design transfers paired supervision through a shared conditional map while retaining modality-specific local tasks.

The resulting learning process connects three quantities: the correspondence of a queried feature, the representation gradient it induces, and retrieval after federated aggregation. We analyze these connections and evaluate them through system experiments, controlled generator comparisons, and query-path substitutions. The controls include a deterministic translator, allowing us to examine the effect of output guidance across architectures. The federation studies then measure how prior learning and its local use work together as client composition and participation change.

Our contributions are: **(1)** a federated conditional-prior protocol that couples guidance through the deployed query with paired-only prior learning and detached local use; **(2)** an analysis of how guidance changes normalized query representations and how auxiliary gradients interact with a receiving client's real objective; and **(3)** experiments on Flickr30k and MS-COCO that connect federation performance with component ablations, generator controls, query sensitivity, and generation cost.

2 Related Work

Multimodal federated learning. Federated optimization combines local training with averaging, proximal regularization, drift correction, or representation contrast (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2020; Li et al., 2021). Multimodal federation additionally varies modality access (Che et al., 2023; Feng et al., 2023). Existing methods coordinate joint representations (Xiong et al., 2022; Zhao et al., 2022), client correlations through FedMSplit (Chen & Zhang, 2022), and modality-specific contributions through FedMobile (Liu et al., 2025).

Representation and structural transfer. CreamFL combines local contrastive objectives with representation ensembles on a shared paired corpus (Yu et al., 2023). FedCola instead develops complementary training and collaborative aggregation for multimodal transformers (Sun et al., 2024). FedTaste transfers relations among semantic groups through a global structural representation and modality-adaptive prompts (Liang et al., 2026). Its transfer target is group structure, whereas FedCGP produces an observation-conditioned feature for local regularization. Both separate the contribution of paired clients from the information available at incomplete clients, but communicate and apply that information differently. FedCGP uses the transferred feature to regularize the locally observed representation.

Knowledge distillation and missing-feature prediction. FedMD transfers predictions through shared data (Li & Wang, 2019); FedGen learns a generator for local regularization without a proxy dataset (Zhu et al., 2021). Multimodal generative models can also exploit partially observed examples (Wu & Goodman, 2018). Client-adaptive reconstruction and feature imputation predict unavailable views from observed modalities (Xiong et al., 2023; Poudel et al., 2025). P-FIN estimates imputation uncertainty and uses it for local gating and aggregation (Shahid et al., 2026). FedCGP instead regulates the flow of supervision: real pairs train the prior, guidance shapes its queried outputs, and incomplete clients use detached predictions to train an observed encoder. Our deterministic controls study output guidance through an MLP translator alongside the denoising prior.

Generative priors and efficient querying. Diffusion models predict noise on corrupted observations (Sohl-Dickstein et al., 2015; Ho et al., 2020); latent diffusion moves this process into a learned representation space (Rombach et al., 2022), and FedDiff explores diffusion in multimodal federation (Li et al., 2024). Efficient generation depends on how a model is trained and queried: DDIM changes the sampling trajectory, consistency models learn maps that support direct generation, and score identity distillation trains a one-step generator (Song et al., 2021; 2023; Zhou et al., 2024). Our Gaussian-seeded map serves as an auxiliary representation, with guidance applied through the operation used by incomplete clients. This objective targets cross-modal representation learning rather than distributional sampling. The fixed-checkpoint study examines whether the learned representation transfers to a different query procedure.

3 FedCGP Framework

3.1 Problem Definition and Overview

Client k observes modalities $\mathcal{M}_k \subseteq \{v, t\}$ and retains its data $\mathcal{D}_k$ locally. We consider client-level modality incompleteness: paired clients $\mathcal{P}$ possess genuine image–text associations, whereas $\mathcal{U}_v$ and $\mathcal{U}_t$ natively observe only one modality and lack paired supervision for the complementary view. An incomplete client also has an observed-data objective $\mathcal{L}_{\text{obs},k}$ and task head $H_{\psi_{\tau(k)}}$. In our federation, image-only and text-only clients perform supervised classification on their own labeled data, while paired clients learn image–text retrieval. Section 5.1 describes the federation and generator experiments.

Encoders E_v, E_t and their projections produce a common d-dimensional representation. Clients aggregate only compatible blocks of the same modality and architecture; task heads additionally require a common task identity and label vocabulary. The bidirectional prior has parameters $\phi = (\phi_{v\to t}, \phi_{t\to v})$. Its outputs are training targets. Final classification uses observed inputs, and retrieval compares real image and text features. Clients communicate model parameters; the generated auxiliary features are computed and consumed locally.

Figure 2 summarizes a round. Paired clients learn real correspondence and update the prior through conditional fitting and single-step guidance (Section 3.2). Incomplete clients use the received prior as a fixed source of detached targets (Section 3.3). The server then synchronizes the supported parameter blocks (Section 3.4). Algorithm 1 and Table A1 in Appendix A detail the execution and update permissions.

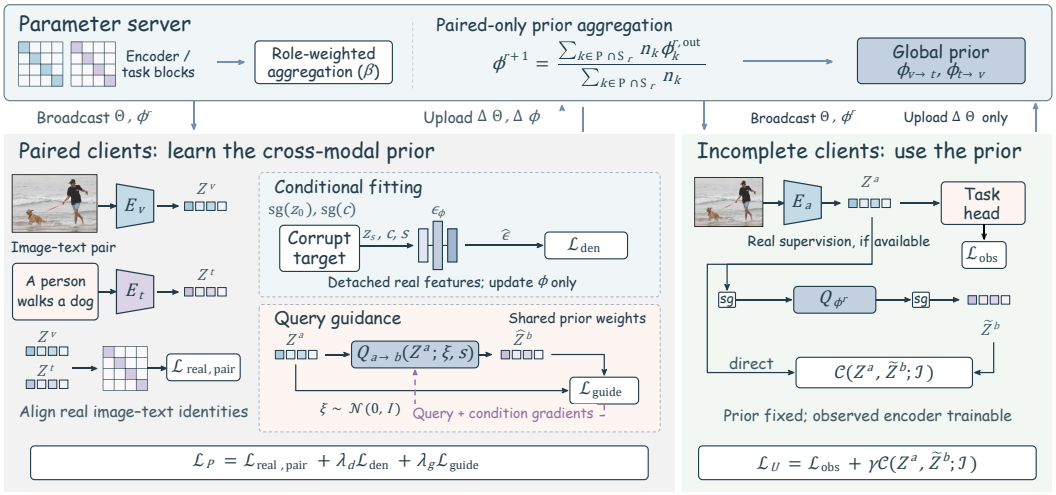


Figure 2: Overview of FedCGP. Paired clients learn the prior; incomplete clients use it with detached targets. The image-only case is shown; text-only use is symmetric.

3.2 Learning the Prior from Paired Observations

Real-pair alignment. For a genuine pair (x^a, x^b), let $z^a = N_\varepsilon(E_a(x^a))$ and $z^b = N_\varepsilon(E_b(x^b))$, where $N_\varepsilon(z) = z/\max(\|z\|_2, \varepsilon)$. For batches of normalized features A, B, define the contrastive objective (Chen et al., 2020; Radford et al., 2021)

$$\mathcal{C}(A, B; P) = -\frac{1}{|A|}\sum_i \log \frac{\sum_{j\in P(i)} \exp(a_i^\top b_j/\tau)}{\sum_j \exp(a_i^\top b_j/\tau)}. \tag{1}$$

Real image–caption identities define $P(i)$. Averaging both real retrieval directions gives $\mathcal{L}_{\text{real,pair}}$. With distinct identities, $P(i) = \{i\}$; repeated identities share positive sets rather than becoming false negatives. The same identity convention applies to generated pairs.

Conditional denoising. For direction $a \to b$, let $c = z^a$ and $z_0 = z^b$. The fitting branch uses stop-gradient copies of both features. A schedule $\bar\alpha_s = \prod_{j=1}^{s}(1-\delta_j)$ defines

$$z_s = \sqrt{\bar\alpha_s}\,\mathrm{sg}(z_0) + \sqrt{1-\bar\alpha_s}\epsilon, \qquad \mathcal{L}_{\text{den}}^{a\to b} = \tfrac{1}{d}\mathbb{E}_{s,\epsilon}\|\epsilon - \epsilon_{\phi_{a\to b}}(z_s, s, \mathrm{sg}(c))\|_2^2, \quad \epsilon\sim\mathcal{N}(0,I). \tag{2}$$

The timestep-conditioned residual MLP predicts injected noise (Ho et al., 2020). This branch updates the prior, not its current conditioning and target encoders. We sum the two directional mean-squared errors to obtain $\mathcal{L}_{\text{den}}$.

Guiding the queried representation. The incomplete client cannot form z_s from an absent real target. We instead parameterize its output by an independent Gaussian state:

$$u_\phi(c;\xi,s) = \frac{\xi - \sqrt{1-\bar\alpha_s}\epsilon_\phi(\xi,s,c)}{\sqrt{\bar\alpha_s}}, \qquad Q_\phi(c;\xi,s) = N_\varepsilon(u_\phi), \quad \xi\sim\mathcal{N}(0,I). \tag{3}$$

This map requires one directional denoiser evaluation, with $T = 1{,}000$ and $s \sim \mathrm{Unif}\{200,\dots,799\}$. Guidance trains its output on independent Gaussian inputs, while conditional denoising regularizes the shared network on corrupted real targets.

On paired clients, write $\widehat{Z}_t = Q_{v\to t}(Z_v)$ and $\widehat{Z}_v = Q_{t\to v}(Z_t)$. Guidance and the complete paired objective are

$$\mathcal{L}_{\text{guide}} = \tfrac{1}{2}\{\mathcal{C}(Z_v, \widehat{Z}_t; \mathcal{I}) + \mathcal{C}(Z_t, \widehat{Z}_v; \mathcal{I})\}, \tag{4}$$

$$\mathcal{L}_P = \mathcal{L}_{\text{real,pair}} + \lambda_d\mathcal{L}_{\text{den}} + \lambda_g\mathcal{L}_{\text{guide}}. \tag{5}$$

Here $\mathcal{I}$ links observations to their conditioned outputs, with shared positives for repeated identities. Guidance differentiates through both the query network and the observed representation. Real-pair alignment anchors the common space, while denoising and query guidance train two branches with shared weights (Figure 3). We use $\lambda_d = \lambda_g = \lambda_{\text{sys}} a_r$, where a_r ramps to one during prior warm-up (Appendix A). Section 5.3 introduces a target-space control for the frozen-feature study.

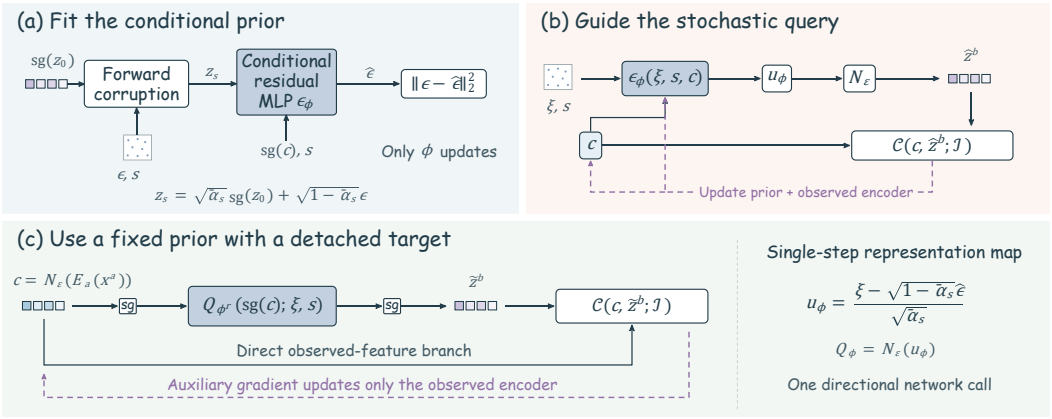


Figure 3: Conditional fitting, query guidance, and detached use. Gradients follow the client-specific paths shown; the network blocks are schematic.

3.3 Prior-Guided Learning at Incomplete Clients

An incomplete client observes Z^a but not Z^b. It queries the received prior without updating it:

$$\widetilde{Z}^b = \mathrm{sg}\big(Q_{\phi^r_{a\to b}}(\mathrm{sg}(Z^a);\xi,s)\big)\,, \qquad \mathcal{L}_{U,k} = \mathcal{L}_{\mathrm{obs},k} + \gamma_r \mathcal{C}(Z^a, \widetilde{Z}^b; \mathcal{I}). \tag{6}$$

The auxiliary term updates the observed representation. A task head receives gradients from its actual observed-data objective, for example cross-entropy on a labeled classification task. The absent-modality encoder and the prior receive no update. The system activates this auxiliary term after prior warm-up, with $\gamma_r = \lambda_{\mathrm{sys}}/2$ thereafter. This ties its strength to the paired-client learning coefficient while preserving the real classification objective.

Fixing ϕ alone would still allow gradients through the prior to its input. Detaching its output removes this indirect path while retaining the direct gradient through Z^a in $\mathcal{C}$. The generated target shapes that gradient; its alignment with the real objective determines the receiving-task effect. Evaluation uses real observations.

3.4 Federated Optimization

Let $\mathcal{S}_r$ denote the participants in round r, and let $\mathcal{S}_b \subseteq \mathcal{S}_r$ contain those that update a compatible representation or task block Θ_b. Image and text blocks use their observing clients; a task head uses clients with the same task identity and label vocabulary. Set $\omega_k = \beta$ for paired clients and $\omega_k = 1$ otherwise, with $\beta > 0$. The server updates

$$\Theta_b^{r+1} = \frac{\sum_{k\in\mathcal{S}_b} n_k \omega_k \Theta_{k,b}^{r,\mathrm{out}}}{\sum_{k\in\mathcal{S}_b} n_k \omega_k}, \qquad \phi^{r+1} = \frac{\sum_{k\in\mathcal{P}\cap\mathcal{S}_r} n_k \phi_k^{r,\mathrm{out}}}{\sum_{k\in\mathcal{P}\cap\mathcal{S}_r} n_k}. \tag{7}$$

Here n_k is the local training count. Only paired clients upload prior parameters, so a common paired weight cancels in their normalized average. In contrast, β changes the balance between paired and incomplete contributors to shared representation blocks. It expresses the role of paired supervision, rather than a client-quality score. An empty contributor set retains its parameter block.

Each round broadcasts compatible blocks, performs the corresponding local updates, and aggregates eligible contributions. The prior is fixed at incomplete clients and generated features stay local. Appendix A gives the procedure and update ownership.

4 Analysis of Learning Guidance

For fixed c, ξ, s, set $v_\phi = \xi - \sqrt{1-\bar{\alpha}_s}\epsilon_\phi(\xi, s, c)$. Where the normalization floor is inactive, $q = Q_\phi = v_\phi/\|v_\phi\|$ because the positive factor $1/\sqrt{\bar{\alpha}_s}$ cancels. With denoiser Jacobian J_ϕ, the gradient of a loss on q is

$$\nabla_\phi \ell(q) = -\frac{\sqrt{1-\bar{\alpha}_s}}{\|v_\phi\|} J_\phi^\top (I - qq^\top)\nabla_q \ell. \tag{8}$$

The projection acts tangentially to the normalized representation. Guidance can therefore change its direction and cross-modal scores even when scale is removed. Appendix B gives the normalization-floor case and full derivation.

Proposition 1 (Local effect of joint guidance)**.** *Fix representations and randomness. Let $f(\phi)$ be the guidance loss, $d(\phi)$ the denoising loss, $g = \nabla f$, and $h = \nabla d$. If f is L-smooth on the update segment, the step $\phi^+ = \phi - \eta(\lambda_g g + \lambda_d h)$ satisfies*

$$f(\phi^+) \le f(\phi) - \eta\big(\lambda_g \|g\|^2 + \lambda_d \langle g, h\rangle\big) + \frac{L\eta^2}{2}\|\lambda_g g + \lambda_d h\|^2. \tag{9}$$

Denoising affects this learning objective through gradient alignment, whereas guidance contributes the explicit term $\lambda_g\|g\|^2$. Their alignment and the step size determine the change in the guidance objective.

Receiving-side effect. For a fixed round, write $\gamma = \gamma_r$. A generated target affects the real task through the representation update it induces. Let g_0 and b be the gradients of the real and auxiliary support losses on the same trainable representation parameters. For a real query loss F with L_F-Lipschitz gradient on both update segments, the comparison between $w_0^+ = w - \eta g_0$ and $w_1^+ = w - \eta(g_0 + \gamma b)$ gives

$$F(w_0^+) - F(w_1^+) = \eta\gamma\langle \nabla F(w), b\rangle + E, \quad |E| \le \tfrac{1}{2} L_F \eta^2 \big(\|g_0\|^2 + \|g_0 + \gamma b\|^2\big). \tag{10}$$

Thus generated-to-real correspondence and receiving-task improvement measure different effects: the latter also depends on the auxiliary gradient's alignment with the real objective. The full derivation is in Appendix B.

5 Experiments

5.1 Experimental Setup

We study federation-level image–text retrieval, generated-to-real correspondence, and the cost of querying the learned prior. System experiments train 384-dimensional MoME-small representations; frozen-feature controls examine the conditional map independently of encoder training.

System experiments. Following the client-role construction and retrieval evaluation of FedCola (Sun et al., 2024), we evaluate FedCGP on Flickr30k and MS-COCO (Plummer et al., 2015; Chen et al., 2015) with 32 clients: 8 paired, 12 image-only, and 12 text-only. Image-only clients use CIFAR100 labels, text-only clients use AG_NEWS labels, and paired clients use the retrieval dataset. The Dirichlet coefficient α controls class imbalance in the unimodal partitions. Paired data are partitioned by image identity, keeping associated captions together. Participation experiments vary overall or modality-specific selection rates. Baseline results are taken from FedCola (Sun et al., 2024) and FedTaste (Liang et al., 2026). FedCGP uses 30 rounds, 12 local epochs at every client, and batch 128; component ablations keep this configuration fixed. Appendix C gives implementation details, and Appendix E summarizes the reference configurations.

Metrics and generator controls. Main tables report S_{comp}, the sum of image-to-text and text-to-image R@1 under the 1k-gallery and full-gallery evaluations. Appendix E defines these evaluations and gives the four directional entries for each summary. Generator controls use frozen MoME features, an image-disjoint fit/development split, and three training seeds. Their C@k metric ranks real target features for generated queries, with same-image positives. Appendix C specifies the data, metric, and training budget.

5.2 Retrieval across Federation Settings

Tables 1 and 2 summarize statistical and modality-specific participation settings. Each dataset entry maps to Tables A7 and A8. Here $\mathbf{r} = (r_v, r_t, r_P)$ denotes image-only, text-only, and paired participation rates. FedCGP achieves the highest composite score in every setting on both datasets. Under the default setting, it reaches 101.79 on Flickr30k and 117.23 on MS-COCO, exceeding

Table 1: Retrieval under heterogeneity and participation changes (S_{comp}).

Method	Default $\alpha = .5,\ r = .25$		More heterogeneity $\alpha = .1,\ r = .25$		Less participation $\alpha = .5,\ r = .125$	
	Flickr	COCO	Flickr	COCO	Flickr	COCO
FedAvg	81.08	95.42	81.70	95.32	64.82	83.91
FedProx	78.55	95.16	76.33	95.62	63.33	79.88
FedIoT	85.51	98.40	83.28	95.89	61.94	80.65
CreamFL	74.83	95.26	80.00	91.41	66.85	78.97
FedCola	91.96	105.10	91.82	100.83	88.85	102.30
FedTaste	99.64	114.31	97.82	109.14	94.68	110.26
FedCGP	**101.79**	**117.23**	**101.04**	**116.25**	**99.18**	**113.95**

Table 2: Retrieval under modality-specific participation (S_{comp}).

Method	More image $\mathbf{r} = (.33, .25, .25)$		More text $\mathbf{r} = (.25, .33, .25)$		Fewer paired $\mathbf{r} = (.25, .25, .125)$	
	Flickr	COCO	Flickr	COCO	Flickr	COCO
FedAvg	78.28	97.28	79.69	96.69	61.12	75.10
FedProx	79.25	95.39	77.59	94.96	59.19	73.08
FedIoT	82.74	95.04	78.02	97.04	60.34	76.14
CreamFL	80.31	93.65	80.75	92.81	57.12	74.69
FedCola	91.24	104.22	90.10	100.96	85.68	94.40
FedTaste	97.15	113.72	97.67	109.42	91.64	102.77
FedCGP	**100.12**	**115.37**	**99.67**	**113.52**	**96.94**	**109.22**

FedCola by 9.83 and 12.13 points. The gains over FedCola range from 8.88 to 11.26 points on Flickr30k and from 11.15 to 15.42 on MS-COCO across the six settings. Compared with FedTaste, the default gains are 2.15 and 2.92 points. Bold values identify the best result within each setting and dataset.

With fewer paired-client updates, FedCGP scores 96.94/109.22 on Flickr30k/MS-COCO, compared with 91.64/102.77 for FedTaste. Reducing paired participation still lowers performance, but the method retains its advantage when paired updates arrive less often. This experiment holds client populations fixed; Figure 4 instead varies the fraction of paired clients.

5.3 Learning and Querying the Prior

Table 3 compares a denoiser and an MLP with and without guidance in the same frozen feature space. The MLP minimizes target-feature MSE; the denoiser minimizes Eq. (2) and is queried through Eq. (3). As a mechanism diagnostic, the guided conditions add

$$\mathcal{L}_{\text{diag}} = \tfrac{1}{2}\{\mathcal{C}(\widehat{Z}_t, Z_t; P) + \mathcal{C}(\widehat{Z}_v, Z_v; P)\}. \tag{11}$$

Eq. (11) aligns generated queries with real targets, while Eq. (4) aligns observations with generated targets during federation. With encoders fixed, this diagnostic isolates the effect of directly supervising the query output. The denoiser and MLP share data and update counts and have 1,185,536 and 394,496 parameters, respectively.

The generated-to-real diagnostic loss raises single-step denoiser C@1 from 0.37/0.35 to 61.75/46.17. This improvement concerns the Gaussian-seeded query, whose inputs differ from corrupted real targets used for denoising. The MLP remains stronger: guided MLP reaches 67.32/51.27, although its image-to-text C@1 decreases slightly from the unguided value. These results support direct supervision of the queried representation. In this controlled feature space, strong correspondence can also be obtained with a deterministic translator.

Table 3: Mechanism diagnostic: architecture and target guidance. C@1/5/10 (%), mean ± SD over three training seeds.

Generator	Target guide	Direction	C@1	C@5	C@10
Denoiser	Off	$v \to t$	0.37 ± 0.16	1.45 ± 0.23	2.92 ± 0.41
Denoiser	Off	$t \to v$	0.35 ± 0.09	0.59 ± 0.19	0.94 ± 0.41
Denoiser	On	$v \to t$	61.75 ± 1.35	87.25 ± 0.76	92.98 ± 0.28
Denoiser	On	$t \to v$	46.17 ± 0.98	55.55 ± 0.78	67.69 ± 1.28
MLP	Off	$v \to t$	$\mathbf{67.91} \pm 1.32$	88.92 ± 0.50	94.98 ± 2.03
MLP	Off	$t \to v$	48.98 ± 1.51	58.20 ± 1.69	70.30 ± 1.24
MLP	On	$v \to t$	67.32 ± 1.29	$\mathbf{89.98} \pm 1.26$	$\mathbf{95.17} \pm 0.89$
MLP	On	$t \to v$	$\mathbf{51.27} \pm 0.31$	$\mathbf{60.91} \pm 0.47$	$\mathbf{72.34} \pm 0.50$

Table 4: Same-checkpoint query paths. C@k (%), mean ± SD over five sampling seeds.

Path	Direction	C@1	C@5	C@10
Single-step	$v \to t$	$\mathbf{52.09} \pm 0.99$	$\mathbf{79.68} \pm 0.61$	$\mathbf{88.08} \pm 0.36$
Single-step	$t \to v$	$\mathbf{39.52} \pm 0.23$	$\mathbf{48.20} \pm 0.49$	$\mathbf{59.28} \pm 0.58$
Reverse-20	$v \to t$	0.87 ± 0.26	3.19 ± 0.54	5.72 ± 0.68
Reverse-20	$t \to v$	1.08 ± 0.19	1.58 ± 0.35	2.51 ± 0.46

The architectures differ in how their fitting losses constrain the deployed query. The MLP predicts a target feature directly, whereas the denoiser predicts injected noise and returns a feature through Eq. (3). Output guidance supervises that returned representation in either case. The comparison therefore examines the learning objective as well as the architecture. Appendix D provides feature proximity and positive-rank diagnostics for these runs.

Query procedure. At a separate fixed checkpoint, single-step C@1 is 52.09/39.52 versus 0.87/1.08 for reverse-20 on the same 400-image cache (Table 4). Five seeds vary sampling at fixed weights. This drop shows the dependence on the query procedure: the checkpoint produces stronger correspondence through its trained single-step interface.

5.4 Component Ablation and Efficiency

Component ablation. Table 5 progressively introduces the conditional prior, query guidance, paired-only prior updates, and paired-source weighting. The Flickr composite score rises from 81.08 to 101.79, with the largest increment from introducing the prior. Each gain is measured with the preceding components in place. This system ablation complements the fixed-feature diagnostic; Appendix F compares deterministic generation.

Table 5: Cumulative ablation.

Configuration	S_{comp}	Δ
No cross-modal prior	81.08	–
+ Conditional prior	90.15	**9.07**
+ Query guidance	94.62	4.47
+ Paired-only prior update	97.05	2.43
+ Paired-source weighting	**101.79**	4.74

Single-step generation takes 0.715 ms at the median versus 15.042 ms for twenty calls (Table 6), an approximately 21-fold reduction. Timing covers the query computation; memory is total process CUDA allocation. Appendix D gives the profiling setup.

Table 6: Latent-generation cost.

	Single-step	Reverse-20
Network calls	**1**	20
Mean (ms)	**0.729**	18.052
Median (ms)	**0.715**	15.042
Peak (MiB)	**38.291**	44.432

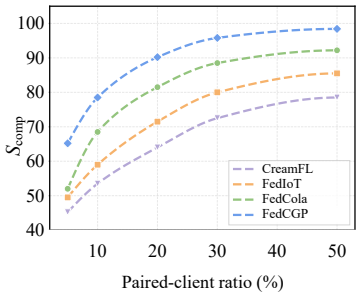


Figure 4: Paired-supervision availability.

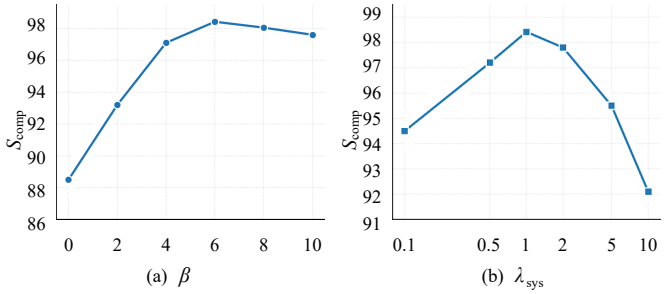


Figure 5: Parameter sensitivity: (a) β; (b) λ_{sys}.

Communication. FedCGP reports 278.07 MB of communication overhead (Table 7). Selected clients receive the prior, and paired contributors upload its updates. The 1,185,536-parameter bidirectional prior occupies 4.522 MiB in FP32 per recipient. Appendix F describes the accounting conventions; Appendix D gives the payload calculation.

Table 7: Reported communication cost.

Method	Overhead
FedAvg	208.81 MB
FedProx	208.81 MB
CreamFL	211.74 MB
FedIoT	208.81 MB
FedCola	371.26 MB
FedTaste	174.58 MB
FedCGP	278.07 MB

Clients receive a prior once per selected round and reuse it across local minibatches. Generated features remain local, so additional queries do not increase the communicated payload. Single-step querying therefore reduces repeated local computation, while paired-only prior updates determine which clients contribute to prior traffic.

5.5 Paired Supervision and Parameter Sensitivity

Figure 4 varies the fraction of paired clients. FedCGP remains above the compared methods throughout the range, with diminishing gains as paired supervision becomes more abundant. This changes the supply of paired clients, complementing the participation study that varies how often a fixed population contributes.

Figure 5(a) peaks near $\beta = 6$, supporting a balance between paired supervision and unimodal task updates in the shared representation blocks. Paired-only prior averaging remains unchanged. Panel (b) peaks near $\lambda_{\text{sys}} = 1$. This coefficient jointly scales denoising, query guidance, and post-warm-up incomplete-client regularization, controlling their strength relative to real-task learning. Appendix F gives the coefficient relations.

Interpreting the component studies. The deterministic-generator system reaches 93.55 versus 101.79 for FedCGP (Appendix F), while the MLP is strong in the frozen-feature study (Table 3). Together, these results show why architecture selection should consider both target correspondence and its interaction with encoder adaptation and federated aggregation.

6 Conclusion

FedCGP learns cross-modal priors from paired clients and supplies detached auxiliary targets to incomplete clients. Guidance trains the same single-step map used for local regularization, while block-specific aggregation separates prior learning from representation sharing. The analysis and experiments show why query correspondence, receiving-side gradients, and federation-level retrieval must be considered together. Controlled feature experiments further show that useful conditional targets can be learned with different architectures. Their value in federation depends on how those targets interact with real supervision and shared representation updates.

AI USE STATEMENT

AI tools were used primarily to improve the language and readability of the manuscript, including grammar correction, wording refinement, conciseness, and stylistic consistency. They were also used for minor checks of notation, formatting, and presentation. All revisions were reviewed by the authors, who take full responsibility for the technical content, experimental results, and final text.

REPRODUCIBILITY STATEMENT

Appendix A details the training procedure, parameter-update ownership, and round-level execution. Appendix B provides the theoretical derivations. Appendices C–F specify the experimental configurations, evaluation metrics, additional results, and profiling setup. We report the distinct sources of variation used in the experiments, including federation configurations, generator-training seeds, stochastic-query seeds, and hardware timing repetitions, so that the scope of each reported result is explicit.

ETHICS STATEMENT

Training keeps raw examples and generated features local while communicating learned parameters. This communication pattern alone does not provide differential privacy or protection against model inversion. Appendix H discusses the threat model, leakage channels, and poisoning risks. Paired-data coverage and access to relevant modalities can differ across participants; applications should consider these differences and the permissions attached to their data and pretrained models.

REFERENCES

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2938–2948, 2020.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32897–32912, 2022.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017. doi: 10.1145/3133956.3133982.

Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. Multimodal federated learning: A survey. *Sensors*, 23(15):6986, 2023. doi: 10.3390/s23156986.

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*, pp. 87–96, 2022. doi: 10.1145/3534678.3539384.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. URL https://arxiv.org/abs/1504.00325.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

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*, pp. 4035–4045, 2023. doi: 10.1145/3580305.3599825.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14 (1–2):1–210, 2021. doi: 10.1561/2200000083.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143, 2020.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015. doi: 10.1109/CVPR.2015.7298932.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Daixun Li, Weiying Xie, Zixuan Wang, Yibing Lu, Yunsong Li, and Leyuan Fang. FedDiff: Diffusion model driven federated learning for multi-modal and multi-clients. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10353–10367, 2024.

Daliang Li and Junpu Wang. FedMD: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019. URL https://arxiv.org/abs/1910.03581.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10713–10722, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Haochen Liang, Jie Zhang, and Hideya Ochiai. FedTaste: Topology-aware structural transfer for multimodal federated learning with missing modalities. *arXiv preprint arXiv:2607.23245*, 2026. URL https://arxiv.org/abs/2607.23245v1.

Yi Liu, Cong Wang, and Xingliang Yuan. FedMobile: Enabling knowledge contribution-aware multi-modal federated learning with incomplete modalities. In *Proceedings of the ACM on Web Conference 2025*, pp. 2775–2786, 2025. doi: 10.1145/3696410.3714623.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, 2017.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2641–2649, 2015.

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*, pp. 289–299. Springer, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Nafis Fuad Shahid, Maroof Ahmed, Md Akib Haider, Saidur Rahman Sagor, Aashnan Rahman, and Md Azam Hossain. Probabilistic feature imputation and uncertainty-aware multimodal federated aggregation. In *Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 3593–3607, 2026.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023.

Guangyu Sun, Matias Mendieta, Aritra Dutta, Xin Li, and Chen Chen. Towards multi-modal transformers in federated learning. In *European Conference on Computer Vision*, pp. 229–246, 2024. doi: 10.1007/978-3-031-72633-0_13.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. A unified framework for multi-modal federated learning. *Neurocomputing*, 480:110–118, 2022. doi: 10.1016/j.neucom.2022.01.063.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1241–1249, 2023. doi: 10.1145/3581783.3611757.

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

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on IoT data. In *IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation*, pp. 43–54, 2022. doi: 10.1109/IoTDI54339.2022.00011.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62307–62331, 2024.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12878–12889, 2021.

APPENDIX

A TRAINING PROTOCOL AND IMPLEMENTATION

A.1 ROUND-LEVEL PROCEDURE

Algorithm 1 specifies one round in the role-pure setting and is repeated for the prescribed training horizon. Each local model starts from the relevant broadcast blocks. Paired clients evaluate the real-pair, denoising, and guidance losses within their local budget. Incomplete clients use a fixed generator and update only their observed representation and compatible task head. The local optimizer acts on the sum of applicable losses, rather than sequentially treating them as independent federated updates.

Algorithm 1 One federated round of FedCGP

Require: Representations, task heads, prior ϕ^r, round r, and local budgets.
1: Set $\lambda_d = \lambda_g = \lambda_{\text{sys}} a_r$ and γ_r using the warm-up schedule.
2: Sample participants and broadcast their compatible parameter blocks.
3: **for** each participating client k in parallel **do**
4: **for** each local batch within the assigned budget **do**
5: **if** client k has paired observations **then**
6: Form detached noisy targets (2) and independent Gaussian queries (3).
7: Update the paired model and prior jointly with Eq. (5).
8: **else**
9: Fix the prior; update the observed representation and head with Eq. (6).
10: **end if**
11: **end for**
12: Return supported blocks and sample counts; only paired clients return the prior.
13: **end for**
14: Aggregate compatible representation blocks and paired prior parameters using Eq. (7).

A.2 DENOISING AND SINGLE-STEP ESTIMATION

The directional denoiser is a timestep-conditioned residual network on latent vectors. Paired fitting uses a real condition and target, with $z_s \sim \mathcal{N}(\sqrt{\bar{\alpha}_s} z_0, (1 - \bar{\alpha}_s) I)$. Single-step estimation instead supplies independent $\xi \sim \mathcal{N}(0, I)$ to Eq. (3). The implementation samples s from $\{200, \ldots, 799\}$ on a 1,000-level schedule. The clean-estimate algebra is related to denoising identities (Efron, 2011; Ho et al., 2020); its posterior interpretation requires the appropriate forward-state distribution. A reverse trajectory repeatedly calls the network under a scheduler and is a different computation. The same single-step map is used for paired guidance and incomplete-client auxiliary targets.

One network evaluation (NFE) counts a call to one directional denoiser. A paired training update can require both directions, both loss branches, and real-task computation. It is not a one-NFE training update. No raw missing image or sentence is used to query Eq. (3).

A.3 GRADIENT PATHS AND TASK SUPERVISION

Table A1 gives the update ownership. Denoising detaches the real condition and target, while the paired guidance path remains differentiable through the generated representation and its condition. The encoders still evolve through real-pair and guidance updates. Fixed encoders are used only in the controlled diagnostic, not as an additional target-network module in the federated method.

For an incomplete client, let $u = \text{sg}(Q_{\phi^r}(\text{sg}(E_a(x)); \xi, s))$. Its direct observed-feature gradient is

$$\nabla_{\theta_a} \ell(E_a(x), u) = J_{E_a}^{\top} \, \partial_{z^a} \ell(z^a, u). \tag{12}$$

The target value changes this derivative even though its own gradient is stopped. A classification head receives its gradient from the real classification objective, not from a feature-level contrast that bypasses the head. Image-only and text-only classification heads are optimized with their respective real labels; retrieval and classification heads are not averaged together.

Table A1: Training signals and parameter updates.

Learning signal	Required local information	Representation	Prior ϕ	Task-specific head
Real-pair retrieval	Image–text identities	Both views	–	–
Conditional denoising	Real condition and target	Detached	Update	–
Paired learning guidance	Condition and generated target	Both views	Update	–
Observed-task loss	Observed input and task labels	Observed view	Fixed	Update
Generated auxiliary loss	Observed input and fixed prior	Observed view	Fixed	–
Diagnostic guidance	Frozen paired feature cache	Fixed	Update	–

Conditioning associations provide the auxiliary positive indices in Eqs. (4) and (6). With repeated image identities, the corresponding positive sets must be shared or the batch must contain distinct identities. Real benchmark evaluation uses real query/gallery observations and their genuine matching identities. The target-space diagnostic objective in Eq. (11) matches generated features to a real target gallery; it is not the same loss as client guidance.

A.4 Loss coupling and activation

The system coupling coefficient λ_{sys} sets the strength of cross-modal prior learning and transfer. With a warm-up horizon $R_w = 10$, define $a_r = \min(r/R_w, 1)$ for rounds $r \geq 1$. The paired objective uses $\lambda_d = \lambda_g = \lambda_{\text{sys}} a_r$. Incomplete clients use $\gamma_r = 0$ for $r \leq R_w$ and $\gamma_r = \lambda_{\text{sys}}/2$ thereafter. Thus the prior first learns from real pairs before supplying auxiliary targets to incomplete clients. The default final coefficient is $\lambda_{\text{sys}} = 1$; the sensitivity sweep varies this value with the activation schedule fixed.

The directional denoising terms average over examples and feature coordinates, and their two directions are summed. The real-pair and guidance terms average their two directions. These reductions fix the meaning of λ_{sys}; changing a sum to an average would change its numerical scale.

A.5 Aggregation and inactive blocks

For each parameter block, the aggregation denominator includes only eligible clients that actually return that block. Define $n_{P,b} = \sum_{k \in \mathcal{P} \cap \mathcal{S}_b} n_k$ and $n_{U,b} = \sum_{k \in \mathcal{U} \cap \mathcal{S}_b} n_k$. When both groups contribute, the representation update can be written as

$$\Theta_b^{r+1} = \rho_b(\beta)\overline{\Theta}_{P,b} + (1 - \rho_b(\beta))\overline{\Theta}_{U,b}, \qquad \rho_b(\beta) = \frac{\beta n_{P,b}}{\beta n_{P,b} + n_{U,b}}, \tag{13}$$

where each group mean is weighted by its local sample counts. For the prior, incomplete clients return no parameter block. Consequently,

$$\frac{\sum_{k \in \mathcal{P} \cap \mathcal{S}_r} \beta n_k \phi_k^{r,\text{out}}}{\sum_{k \in \mathcal{P} \cap \mathcal{S}_r} \beta n_k} = \frac{\sum_{k \in \mathcal{P} \cap \mathcal{S}_r} n_k \phi_k^{r,\text{out}}}{\sum_{k \in \mathcal{P} \cap \mathcal{S}_r} n_k}, \quad \beta > 0. \tag{14}$$

Equation (7) therefore uses paired-only normalization for the prior. No unchanged incomplete-client copy enters its denominator. If there is no eligible contributor with positive weight, the server retains the old block. At the zero-weight endpoint $\beta = 0$, paired contributions are excluded, including prior updates; positive-β settings preserve paired-only prior averaging. Without paired data at initialization, training reduces to the observed tasks until paired supervision becomes available.

A.6 Mixed availability

A mixed client could fit directional priors on its genuinely paired examples and use them on incomplete examples. Such an extension requires subset-specific data counts and update eligibility. The experiments here use role-pure client groups. Additional modality directions also require observed condition–target support; they do not follow solely from matching feature dimensions.

B Mechanism Analysis and Proofs

B.1 Why paired observations contain distinct information

Let X be uniform on $\{-1, +1\}$. Consider P_+ with $Y = X$ and P_- with $Y = -X$. Both have uniform marginal distributions for X and Y. A learner receiving independent samples from only those marginals observes identical information under the two worlds, although $P_+(Y \mid X)$ and $P_-(Y \mid X)$ differ everywhere. Paired samples identify the sign relationship. This is an identifiability example in the absence of additional structural assumptions, not a claim that all unpaired learning is uninformative.

The same observation bounds what synthetic targets establish. Sampling from a learned model cannot, by itself, reveal which of two observationally indistinguishable missing-side conditionals is correct. A generated target may still be useful for a local objective, but usefulness is weaker than recovery of the actual missing view.

B.2 Normalized single-step Jacobian

For fixed c, ξ, s, abbreviate $a = \bar{\alpha}_s > 0$, $b = \sqrt{1-a}$, $e_\phi = \epsilon_\phi(\xi, s, c)$, and $v = \xi - be_\phi$. The map is $q = N_\varepsilon(v/\sqrt{a})$. When $\|v\| > \varepsilon\sqrt{a}$,

$$q = \frac{v}{\|v\|}, \qquad \frac{\partial q}{\partial v} = \frac{I - qq^\top}{\|v\|}, \qquad \frac{\partial v}{\partial \phi} = -bJ_\phi. \tag{15}$$

Combining these identities gives Eq. (8). The factor $1/\sqrt{a}$ in the unnormalized clean-estimate parameterization cancels under exact radial normalization. The Jacobian still depends on $\|v\|$, the network Jacobian, and the task gradient, so cancellation is not a universal stability guarantee.

When $\|v\| < \varepsilon\sqrt{a}$, the floor is active and

$$q = \frac{v}{\varepsilon\sqrt{a}}, \qquad \frac{\partial q}{\partial \phi} = -\frac{\sqrt{1-a}}{\varepsilon\sqrt{a}} J_\phi. \tag{16}$$

At the boundary, the map is continuous and piecewise differentiable; the analysis can be interpreted away from the boundary or through an admissible generalized derivative.

B.3 Proof of Proposition 1

Let $D = \lambda_g g + \lambda_d h$. By smoothness on the line segment from ϕ to $\phi - \eta D$,

$$f(\phi - \eta D) \le f(\phi) - \eta\langle \nabla f(\phi), D\rangle + \tfrac{1}{2}L\eta^2\|D\|^2 \tag{17}$$

$$= f(\phi) - \eta\big(\lambda_g\|g\|^2 + \lambda_d\langle g, h\rangle\big) + \tfrac{1}{2}L\eta^2\|D\|^2. \tag{18}$$

This proves Eq. (9). If $D \neq 0$ and the first-order quantity $A = \lambda_g\|g\|^2 + \lambda_d\langle g, h\rangle$ is positive, then $0 < \eta < 2A/(L\|D\|^2)$ is sufficient for the bound to certify descent. If $A \le 0$, the statement does not certify descent. In particular, adding guidance does not guarantee that every stochastic step, every receiver, or the final aggregated task model improves.

B.4 What an auxiliary target changes in local learning

Let g_0 be the gradient of a real support objective and b the gradient of the auxiliary term on the same representation parameters. Compare

$$w_0^+ = w - \eta g_0, \qquad w_1^+ = w - \eta(g_0 + \gamma b). \tag{19}$$

For a real query loss F with L_F-Lipschitz gradient on both update segments, write

$$F(w_j^+) = F(w) + \langle \nabla F(w), w_j^+ - w\rangle + R_j, \qquad |R_j| \le \tfrac{1}{2}L_F\|w_j^+ - w\|^2.$$

Subtracting these two expansions gives Eq. (10), with $E = R_0 - R_1$ and $|E| \le |R_0| + |R_1|$. It clarifies why good feature correspondence need not entail a positive classification gain: the induced auxiliary gradient must also be useful for the real objective on the trainable block. A head outside the auxiliary path receives no such gradient.

Table A2: Full-federation configuration.

Item	Setting
Client population	32: 8 paired, 12 image-only, 12 text-only
Training	30 rounds; 12 local epochs at every client; batch 128
Optimization	AdamW; learning rates 10^{-4} (encoders/heads), 10^{-3} (prior); .99 decay per round; weight decay 0
Representations	MoME-small; shared dimension 384
Denoiser	Bidirectional; two residual blocks per direction, width 256, dropout .1; 1,000 noise levels
Schedule	Linear variance from 10^{-4} to .02
Default participation	$r = .25$; lower-participation setting $r = .125$
Partition settings	$\alpha = .5$ and $\alpha = .1$
Paired-source weighting	$\beta = 6$; role-weighted shared blocks, paired-only prior
Loss coupling	$\lambda_{\text{sys}} = 1$; 10-round prior warm-up; $\gamma_r = \lambda_{\text{sys}}/2$ afterward
Retrieval evaluation	1k-gallery and full-gallery; Appendix E

B.5 Fitting and querying sample different latent states

Denoising is optimized on $z_s \sim \mathcal{N}(\sqrt{\bar{\alpha}_s} z_0, (1 - \bar{\alpha}_s) I)$, whereas the single-step readout uses $\xi \sim \mathcal{N}(0, I)$ at an interior s. These laws generally differ even though both have full support. Thus the issue is a change of input distribution and objective, not necessarily disjoint support. A fitting-only denoiser's weak single-step correspondence cannot be generalized to its fully optimized reverse-generation behavior.

C Experiment Configurations and Metrics

C.1 Federation-level configuration

Table A2 collects the complete-system settings. The full-system experiments use MoME-small with 384-dimensional representations (Bao et al., 2022; Sun et al., 2024). The backbone uses ViT-Small transformer blocks (Dosovitskiy et al., 2021), ImageNet image initialization (Deng et al., 2009), and a BERT tokenizer (Devlin et al., 2019). Local optimization uses AdamW (Loshchilov & Hutter, 2019). Paired clients supply cross-modal supervision, while incomplete clients contribute through observed representations. The observed-task term in Eq. (6) is cross-entropy on CIFAR100 or AG_NEWS labels, respectively. The generator diagnostic instead fixes both encoders to isolate conditional feature transfer. The client population combines 12 CIFAR100 image classifiers, 12 AG_NEWS text classifiers, and eight paired retrieval clients. The epoch budget is identical across client roles. System contrastive terms use temperature $\tau = .05$; the frozen-feature diagnostic uses .07. Class-label Dirichlet partitions use $\alpha = .5$ by default and .1 for stronger unimodal heterogeneity. Paired partitions shuffle image identities, retain their associated captions together, and allow small size imbalance. Retrieval uses the 1k-gallery and full-gallery modes defined in Appendix E; it is distinct from local unimodal classification.

C.2 Frozen representation cache

The controlled study uses a frozen MoME-small reference with 384-dimensional outputs, from the modality-expert transformer family (Bao et al., 2022; Sun et al., 2024). Its cache adapter uses a seed-42 configuration with four image, four text, and two paired clients. From the paired partitions, it retains 8,251 image–caption rows for 2,000 unique images. An 80/20 split by image ID assigns 1,600 identities to generator fitting and 400 identities (1,634 rows) to evaluation. Captions of the same image remain together.

The image-ID split evaluates conditional mapping to held-out identities in a fixed pretrained feature space. Budget selection compared 2,000, 3,000, and 5,000 updates on this development split; the four-condition diagnostic uses 3,000 updates.

Table A3: Controlled generator configuration.

Item	Setting
Reference	Frozen MoME-small, $d = 384$
Data	2,000 images; 8,251 pair-indexed rows
Fit / development	1,600 / 400 images; split seed 42
Training	3,000 AdamW updates; learning rate 10^{-3}; batch 64
Training seeds	1, 2, 3; evaluation seed 12345
MLP	Two $384 \to 256 \to 384$ translators with GELU; 394,496 parameters
Denoiser	Two directional, two-block residual networks; width 256; dropout .1; 1,185,536 parameters
Guidance	Target-space multi-positive contrast; temperature .07
Environment	RTX 4090; Python 3.8.20; PyTorch 2.0.0+cu117

C.3 Objectives of the four conditions

The deterministic translator minimizes

$$\mathcal{L}_{\text{fit}}^{\text{MLP}} = \text{MSE}(M_{v\to t}(Z_v), Z_t) + \text{MSE}(M_{t\to v}(Z_t), Z_v). \tag{20}$$

The denoiser uses Eq. (2); both of its conditions are evaluated through the single-step estimator. Guidance adds the generated-to-real objective $\mathcal{L}_{\text{diag}}$ in Eq. (11). It uses genuine target-modality features and image-ID positives. This diagnostic loss differs from the observed-to-generated paired-client objective $\mathcal{L}_{\text{guide}}$, although both act through the same estimator. Data and update count are held fixed; Table A3 specifies the architecture sizes and training settings.

C.4 Correspondence, proximity, and rank

For generated query row i, rank real target rows by cosine similarity, define $P(i) = \{j : \text{id}(j) = \text{id}(i)\}$, and let r_i be its best positive rank. Then

$$\text{C@}k = \frac{100}{N_q}\sum_{i=1}^{N_q} \mathbf{1}[r_i \le k], \qquad \overline{r} = \frac{1}{N_q}\sum_i r_i. \tag{21}$$

The $v \to t$ direction ranks real text features for generated text queries; $t \to v$ ranks real image features for generated image queries. The cache is pair-indexed: image candidates can repeat and queries are row-weighted. All rows with a matching image identity count as positives. This definition accounts for repeated rows when computing top-k correspondence.

Cosine measures proximity to the associated target row. Raw MSE is computed before metric normalization, so it reflects output scale as well as alignment: denoiser outputs have unit norm, whereas MLP outputs can vary in norm. For reference, independent isotropic unit vectors in 384 dimensions have expected per-coordinate squared distance $2/384 \simeq .0052$.

C.5 Different units of repetition

The factorial study reports variability across three independently initialized generator fits. The query-path study uses five sampling seeds at fixed weights, and profiling uses fifty timing repetitions after ten warm-ups. Means and standard deviations refer to these respective units; federation tables report point estimates.

D Additional Diagnostic Results

D.1 Feature quality and ranking

Table 3 contains all C@1/5/10 values for the four generator conditions. Table A4 adds paired cosine, raw-output MSE, and positive rank for exactly those runs. The guided MLP has lower paired cosine but better mean rank in both directions and higher top-k correspondence in several columns, illustrating the difference between absolute proximity and relative gallery ordering.

Table A4: Feature diagnostics: mean ± SD over three generator fits.

Generator	Objective	Direction	Cosine	Raw MSE	Mean rank
Denoiser	Fitting	$v \to t$	0.010 ± 0.002	0.005 ± 0.000	435.72 ± 15.89
Denoiser	Fitting	$t \to v$	0.008 ± 0.001	0.005 ± 0.000	754.52 ± 7.85
Denoiser	Guided	$v \to t$	0.334 ± 0.002	0.003 ± 0.000	3.87 ± 0.02
Denoiser	Guided	$t \to v$	0.334 ± 0.002	0.003 ± 0.000	25.64 ± 0.13
MLP	Fitting	$v \to t$	$\mathbf{0.554} \pm 0.004$	$\mathbf{0.002} \pm 0.000$	3.05 ± 0.33
MLP	Fitting	$t \to v$	$\mathbf{0.570} \pm 0.002$	$\mathbf{0.002} \pm 0.000$	23.17 ± 0.48
MLP	Guided	$v \to t$	0.527 ± 0.001	0.007 ± 0.000	$\mathbf{2.98} \pm 0.35$
MLP	Guided	$t \to v$	0.538 ± 0.001	0.006 ± 0.000	$\mathbf{19.01} \pm 0.58$

Table A5: Single-step and reverse-20 queries: mean ± SD over five sampling seeds.

Path	Direction	C@1	C@5	C@10	Mean rank
Single-step	$v \to t$	$\mathbf{52.09} \pm 0.99$	$\mathbf{79.68} \pm 0.61$	$\mathbf{88.08} \pm 0.36$	$\mathbf{6.98} \pm 0.18$
Single-step	$t \to v$	$\mathbf{39.52} \pm 0.23$	$\mathbf{48.20} \pm 0.49$	$\mathbf{59.28} \pm 0.58$	$\mathbf{39.52} \pm 1.22$
Reverse-20	$v \to t$	0.87 ± 0.26	3.19 ± 0.54	5.72 ± 0.68	319.22 ± 8.26
Reverse-20	$t \to v$	1.08 ± 0.19	1.58 ± 0.35	2.51 ± 0.46	577.86 ± 13.39

Table A6: Latent-generation profile: batch 32, 50 repetitions after 10 warm-ups.

Path	NFE	Mean (ms)	Median (ms)	Peak allocation (MiB)
Single-step	**1**	**0.729**	**0.715**	**38.291**
Reverse-20	20	18.052	15.042	44.432

For the MLP, guidance lowers paired cosine from 0.554/0.570 to 0.527/0.538, while mean positive rank improves from 3.05/23.17 to 2.98/19.01 ($v \to t/t \to v$; full statistics in Table A4). Pointwise proximity and ranking among competing targets need not agree. Raw MSE also depends on output norms, which differ between the MLP and normalized denoiser.

D.2 Complete query-path diagnostic

This experiment uses a separate Flickr30k generator checkpoint from the factorial comparison. It is evaluated on the 400-image, 1,634-row diagnostic cache, with seeds 1–5. The reverse path applies a 20-step scheduler to the same weights. Its comparison with the single-step query jointly reflects the learned objective, queried-state distribution, and schedule.

D.3 Latency, memory, and payload

Timing uses cached 384-dimensional inputs on an RTX 4090 and excludes raw-input encoding, data loading, communication, and optimizer work. The reported median ratio is 21.02 using underlying measurements; the rounded milliseconds are reported in the table and the paper describes the ratio as approximately 21-fold. Each directional single-step query makes one denoiser call; joint bidirectional training requires more work.

The two-direction generator contains 1,185,536 parameters. FP32 tensor values occupy

$$1{,}185{,}536 \times 4 = 4{,}742{,}144 \text{ bytes} = 4.5224609375 \text{ MiB}. \tag{22}$$

This excludes serialization metadata, framing, encoder parameters, optimizer states, multiple recipients, and repeated rounds. The peak CUDA figures include resident tensors. Neither quantity is a full-federation communication or training-cost estimate.

D.4 Restricted receiver-side pilot

An exploratory CIFAR100/AG_NEWS pilot compared real-task-only updates, detached self-targets, shuffled generated targets, and correctly conditioned targets. With the prior fixed, 100 steps updated only final normalization and classification heads (batch 32, seed 1). Correct conditioning showed no additional accuracy advantage. This local pilot illustrates Eq. (10)'s distinction between target correspondence and auxiliary-gradient utility under restricted adaptation.

E Expanded Federation-Level Results

Tables A7 and A8 provide the directional results underlying the main summaries for all six federation settings. The aggregate score is computed as the sum of the four directional entries for each dataset. Published baseline results follow Sun et al. (2024), and the FedTaste results are taken from Liang et al. (2026).

Following the adopted retrieval protocol, we report 1k-gallery and full-gallery evaluations. The 1k-gallery score averages five disjoint subsets, each containing 1,000 image identities and their associated captions. The full-gallery score uses the complete held-out retrieval pool. Both modes use unique image candidates, retain all associated captions, and share the same encoders and similarity function. Their four-entry summary is

$$S_{\text{comp}} = \text{R@1}_{\text{I2T,1k}} + \text{R@1}_{\text{T2I,1k}} + \text{R@1}_{\text{I2T,full}} + \text{R@1}_{\text{T2I,full}}.$$

This sum has a range of zero to 400 and is measured in score units; each directional recall is a percentage. Each term is a top-1 recall. We use this definition throughout the federation tables, cumulative ablation, and sensitivity plots.

Numerical precision. Directional values are reported to two decimal places. Composite scores retain their reported precision and can differ slightly from the sum of rounded entries. The same default configuration is used in both main-table groups.

Reference configurations. We adopt FedCola's client roles and retrieval evaluation (Sun et al., 2024): CIFAR100 and AG_NEWS classification at unimodal clients (Krizhevsky, 2009; Zhang et al., 2015), paired retrieval, and the six federation settings above. FedCola reports 30 communication rounds with five local epochs; FedCGP uses 30 rounds with 12 local epochs at every client and batch size 128. Table A2 gives the FedCGP architecture and optimization settings. Table 5 keeps the FedCGP training configuration fixed across component additions.

FedAvg and FedProx are per-modality optimization controls (McMahan et al., 2017; Li et al., 2020); FedIoT changes multimodal aggregation (Zhao et al., 2022). The CreamFL adaptation in the reference study distills on 500 public COCO samples and aggregates client models in place of centralized server training (Yu et al., 2023; Sun et al., 2024). FedCola uses complementary training and collaborative aggregation; FedTaste uses frozen encoders, modality-adaptive prompts, and structural transfer. The fixed-cache MLP is a conditional-feature control, not an implementation of a competing federated method.

Benchmark metric. Image–text retrieval evaluates genuine cross-modal correspondences (Karpathy & Fei-Fei, 2015). For real query q, its correct real targets $P(q)$, and top-k candidates $T_k(q)$,

$$\text{R@}k = \frac{100}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[P(q) \cap T_k(q) \neq \varnothing]. \tag{23}$$

Real retrieval uses unique image candidates and treats each matching caption as positive. It differs from C@k in query representation and gallery construction; all federation tables report R@1.

Table A7: Flickr30k directional results (%) and composite score S_{comp}.

Method	1k-gallery I2T	1k-gallery T2I	Full gallery I2T	Full gallery T2I	Composite S_{comp}
Default: $\alpha = .5,\ r = .25$					
FedAvg	32.84	22.90	15.32	10.02	81.08
FedProx	31.36	22.41	14.84	9.94	78.55
CreamFL	30.20	21.34	13.82	9.46	74.83
FedIoT	34.42	23.87	16.34	10.88	85.51
FedCola	35.68	26.14	18.10	12.04	91.96
FedTaste	38.52	28.24	19.68	13.20	99.64
FedCGP	**39.46**	**28.84**	**20.15**	**13.34**	**101.79**
More heterogeneity: $\alpha = .1,\ r = .25$					
FedAvg	32.50	23.34	15.40	10.46	81.70
FedProx	31.10	22.06	13.90	9.26	76.33
CreamFL	31.48	22.59	15.74	10.19	80.00
FedIoT	33.02	23.73	15.94	10.57	83.28
FedCola	36.26	26.06	17.54	11.96	91.82
FedTaste	37.98	27.77	19.14	12.93	97.82
FedCGP	**39.40**	**28.64**	**19.76**	**13.24**	**101.04**
Less participation: $\alpha = .5,\ r = .125$					
FedAvg	25.84	18.95	11.96	8.06	64.82
FedProx	25.94	19.02	10.74	7.64	63.33
CreamFL	26.94	19.54	11.90	8.47	66.85
FedIoT	25.18	18.13	11.02	7.62	61.94
FedCola	34.94	25.48	16.60	11.84	88.85
FedTaste	37.04	26.76	18.43	12.45	94.68
FedCGP	**38.86**	**28.01**	**19.28**	**13.03**	**99.18**
More image: $\mathbf{r} = (.33, .25, .25)$					
FedAvg	31.22	22.68	14.42	9.96	78.28
FedProx	31.46	22.84	14.90	10.05	79.25
CreamFL	32.02	23.18	14.74	10.36	80.31
FedIoT	33.22	23.40	15.78	10.34	82.74
FedCola	35.42	25.80	17.76	12.26	91.24
FedTaste	37.60	27.60	19.08	12.87	97.15
FedCGP	**38.85**	**28.38**	**19.55**	**13.34**	**100.12**
More text: $\mathbf{r} = (.25, .33, .25)$					
FedAvg	31.94	22.55	15.20	10.00	79.69
FedProx	31.20	22.25	14.44	9.70	77.59
CreamFL	31.96	23.20	15.12	10.47	80.75
FedIoT	31.46	22.22	14.56	9.77	78.02
FedCola	35.48	25.50	17.40	11.72	90.10
FedTaste	37.93	27.71	19.06	12.97	97.67
FedCGP	**38.91**	**27.87**	**19.50**	**13.39**	**99.67**
Fewer paired: $\mathbf{r} = (.25, .25, .125)$					
FedAvg	24.92	18.01	10.70	7.49	61.12
FedProx	24.28	17.50	10.22	7.19	59.19
CreamFL	23.20	17.12	9.64	7.15	57.12
FedIoT	24.68	17.76	10.42	7.47	60.34
FedCola	34.06	24.28	16.18	11.16	85.68
FedTaste	36.04	25.94	17.64	12.02	91.64
FedCGP	**38.13**	**27.35**	**18.74**	**12.72**	**96.94**

Table A8: MS-COCO directional results (%) and composite score S_{comp}.

Method	1k-gallery I2T	1k-gallery T2I	Full gallery I2T	Full gallery T2I	Composite S_{comp}
Default: $\alpha = .5,\ r = .25$					
FedAvg	36.98	29.28	16.76	12.40	95.42
FedProx	37.56	28.46	16.68	12.46	95.16
CreamFL	37.60	28.64	16.68	12.34	95.26
FedIoT	38.62	29.97	17.16	12.65	98.40
FedCola	41.02	31.62	18.74	13.72	105.10
FedTaste	44.56	34.44	20.41	14.90	114.31
FedCGP	**45.65**	**35.33**	**20.90**	**15.35**	**117.23**
More heterogeneity: $\alpha = .1,\ r = .25$					
FedAvg	37.46	29.11	16.40	12.35	95.32
FedProx	37.66	28.86	16.90	12.20	95.62
CreamFL	35.76	28.11	15.74	11.80	91.41
FedIoT	37.66	29.47	16.52	12.24	95.89
FedCola	39.62	30.37	17.72	13.12	100.83
FedTaste	42.63	32.89	19.43	14.19	109.14
FedCGP	**45.53**	**34.89**	**20.63**	**15.20**	**116.25**
Less participation: $\alpha = .5,\ r = .125$					
FedAvg	32.68	26.12	14.22	10.90	83.91
FedProx	31.20	25.13	13.28	10.27	79.88
CreamFL	31.58	25.00	12.60	9.79	78.97
FedIoT	31.62	25.35	13.44	10.24	80.65
FedCola	40.12	30.47	18.28	13.43	102.30
FedTaste	43.29	33.12	19.44	14.41	110.26
FedCGP	**44.82**	**34.08**	**20.10**	**14.95**	**113.95**
More image: $\mathbf{r} = (.33, .25, .25)$					
FedAvg	38.26	29.33	17.22	12.47	97.28
FedProx	37.46	28.67	16.80	12.46	95.39
CreamFL	36.80	28.66	16.02	12.17	93.65
FedIoT	36.86	29.06	16.78	12.34	95.04
FedCola	40.58	31.05	19.26	13.33	104.22
FedTaste	44.41	34.26	20.26	14.79	113.72
FedCGP	**45.27**	**34.59**	**20.59**	**14.92**	**115.37**
More text: $\mathbf{r} = (.25, .33, .25)$					
FedAvg	38.00	28.95	17.26	12.48	96.69
FedProx	36.96	28.70	16.92	12.38	94.96
CreamFL	36.68	28.46	15.62	12.06	92.81
FedIoT	37.74	29.47	17.22	12.61	97.04
FedCola	39.82	30.32	17.96	12.86	100.96
FedTaste	42.84	32.95	19.44	14.19	109.42
FedCGP	**44.62**	**33.87**	**20.36**	**14.67**	**113.52**
Fewer paired: $\mathbf{r} = (.25, .25, .125)$					
FedAvg	30.30	23.78	12.02	8.99	75.10
FedProx	29.22	23.32	11.32	9.22	73.08
CreamFL	29.60	23.78	12.18	9.12	74.69
FedIoT	30.80	23.56	12.42	9.36	76.14
FedCola	37.78	28.08	16.80	11.74	94.40
FedTaste	40.51	30.75	18.23	13.28	102.77
FedCGP	**43.31**	**32.40**	**19.61**	**13.90**	**109.22**

F Components and Additional Experiments

F.1 Composite score and cumulative components

The composite score S_{comp} sums the four directional entries in Tables A7 and A8. It is used consistently for the federation summary, cumulative ablation, and sensitivity plots; its evaluation-setting interpretation is given in Appendix E.

The cumulative series is 81.08, 90.15, 94.62, 97.05, and 101.79. Its adjacent gains are 9.07, 4.47, 2.43, and 4.74, for a total of 20.71 score units. Each gain is conditional on all earlier components. The system-level deterministic-generator comparison yields 93.55, versus 101.79 for the complete FedCGP configuration. The fixed-feature factorial evaluates a separate translator with its own architecture and optimization. Together, the studies examine both conditional target quality and its interaction with federated representation learning.

F.2 Paired availability and parameter sensitivity

Figure 4 changes the fraction of clients with paired observations. This differs from changing their participation probability at fixed population composition. Both affect the supply of paired supervision. A population sweep also changes the paired data available to train the prior, so its effect is not attributed solely to the number of contributors.

For a shared representation block, Eq. (13) gives the paired contribution $\rho_b(\beta) = \beta n_{P,b}/(\beta n_{P,b} + n_{U,b})$. If both groups contribute, its derivative is $n_{P,b}n_{U,b}/(\beta n_{P,b} + n_{U,b})^2$. Increasing β therefore increases paired influence with diminishing changes in the mixture. The coefficient does not distinguish quality among paired clients. For the prior, the common positive weight cancels because only paired clients contribute. The zero-weight endpoint excludes those contributions entirely. The score peak near $\beta = 6$ is consistent with a balance between cross-modal alignment and unimodal task updates, rather than a monotonic benefit from emphasizing either group.

The loss sweep varies the final common coefficient λ_{sys}. Paired-client weights satisfy $\lambda_d = \lambda_g = \lambda_{\text{sys}} a_r$, and the incomplete-client weight is $\gamma_r = (\lambda_{\text{sys}}/2)\mathbf{1}\{r > 10\}$. Thus all three terms change together under a fixed schedule. Small λ_{sys} weakens both prior learning and its auxiliary effect on incomplete clients; large λ_{sys} gives these terms greater weight relative to real-task objectives. The maximum near one identifies a useful operating point for this coupled configuration.

F.3 Communication accounting

Table 7 follows the accounting conventions of the corresponding methods: FedCola counts the combined download of one image-only and one text-only client, and FedTaste follows its arXiv report (Liang et al., 2026). FedCGP has a communication overhead of 278.07 MB. Selected clients download the prior, paired contributors upload prior updates, and compatible representation blocks are synchronized separately. The table does not measure total federation traffic over all rounds.

F.4 Qualitative feature views

The heatmap displays pairwise image–text similarity, with the diagonal corresponding to matched observations. The t-SNE projection (van der Maaten & Hinton, 2008) illustrates the arrangement of image and text features in two dimensions. These views complement retrieval scores by showing local cross-modal structure; the projection itself is not a quantitative measure of alignment or generative diversity.

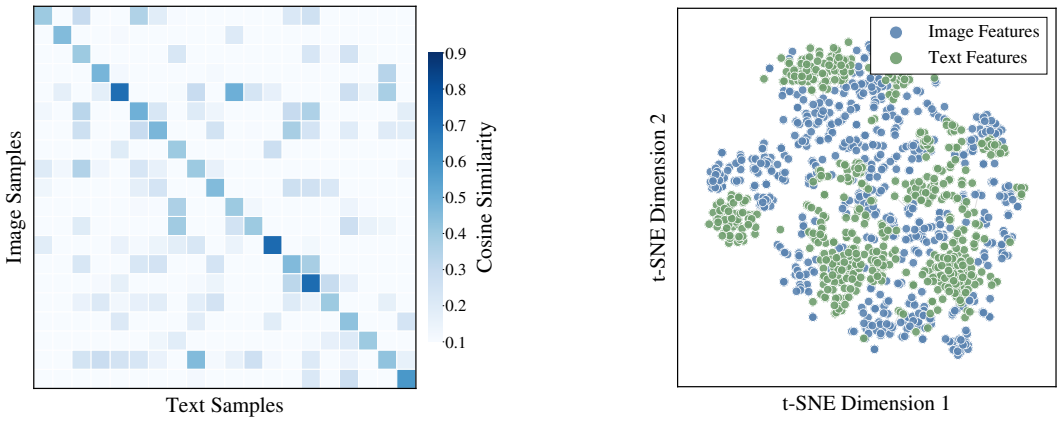


Figure A1: Paired similarity and two-dimensional representation visualization.

G Scope of the Empirical Analysis

The experiments use two complementary levels of evaluation. Federation studies measure retrieval with trainable 384-dimensional MoME-small representations. Generator controls freeze a 384-dimensional reference cache to separate conditional mapping from encoder adaptation. Their three training seeds, five query seeds, and fifty timing repetitions measure different sources of variation and are reported separately.

The framework is evaluated through retrieval, conditional-feature correspondence, and local query cost. The cumulative system ablation includes encoder adaptation and server aggregation; the feature study isolates the conditional map. Conditional diversity and adversarial robustness remain open evaluation directions.

H Security and Privacy Considerations

Threat model and exposure. The protocol assumes authenticated participants follow the prescribed updates. Raw observations and generated features remain local, but the server receives parameter updates, client roles, and sample counts. A curious server or client may infer properties of paired data from those updates or the shared prior. Local data retention alone is not a privacy guarantee; gradient reconstruction provides a concrete leakage mechanism (Zhu et al., 2019).

Limits of isolation. Fixing the prior at incomplete clients prevents their auxiliary loss from directly changing it. Stop-gradient controls optimization, not information disclosure: it cannot remove memorized associations from the prior. Gaussian query noise is not calibrated for differential privacy. Generated targets are model predictions and must not be treated as observed missing modalities.

Poisoning. Malicious paired clients can corrupt conditional relations. Increasing β increases their contribution to shared representations, without assessing trustworthiness or changing relative prior weights. Incomplete clients can poison representation blocks despite exclusion from prior aggregation. Simple averaging remains vulnerable to model-replacement attacks (Bagdasaryan et al., 2020).

Protection requirements. Secure aggregation hides individual updates under its protocol assumptions (Bonawitz et al., 2017), but does not validate them or protect the released prior. Stronger privacy requires explicit clipping, calibrated noise, and composition accounting. Robust aggregation must accommodate modality-dependent non-IID updates. These defenses require separate privacy and attack evaluations; the present protocol provides update isolation rather than a formal security guarantee.