International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 2, Issue 1, 72-79, 2021
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V2I1P109

Original Article

ANCHOR-GEN: Unsupervised Multimodal Generation via Latent Cross-Domain Anchors and Trust-Metric Federated Learning for Integrity, Accountability, and Explainability–Performance Trade-offs

Mohan Siva Krishna Konakanchi
Independent Researcher, USA.

Abstract - *Unsupervised multimodal generation aims to syn- thesize coherent content across domains and modalities (e.g., image and text, audio and video) without relying on paired training data. While recent generative models demonstrate strong perceptual quality, practical deployment across organizational silos remains constrained by privacy, governance, and trust: data cannot be centralized, model updates may be unreliable, and the most accurate models are often least explainable. This paper proposes ANCHOR-GEN, a novel unsupervised multimodal generation framework built around latent cross-domain anchors: compact, interpretable latent factors that align semantically across modalities while remaining learnable from unpaired data. ANCHOR-GEN couples (i) anchor discovery, (ii) anchor- consistent multimodal generation, and (iii) anchor-based explana-tions that expose controllable semantic dimensions of generation. To support deployment across silos, we introduce a trust metric- based federated learning framework that enforces integrity and accountability through update provenance, robust aggregation, privacy-preserving protocols, and auditable trust scoring. Finally, we present a practical framework to quantify and optimize the trade-off between explainability and performance using lightweight metrics based on anchor stability, concept alignment, and utility retention. Experiments on standard unpaired cross- domain settings demonstrate that anchorization improves cross- domain consistency and controllability while enabling measurable explainability with minimal performance loss. The proposed fed- erated trust layer reduces sensitivity to low-quality or adversarial updates and provides accountability without centralizing data.*

Keywords - *Unsupervised generation, multimodal learning, federated learning, trust metrics, accountability, robust aggrega- tion, explainable AI, controllability.*

1. Introduction

Modern enterprises increasingly seek generative AI sys- tems that can operate across heterogeneous data sources and modalities for example, synthesizing product imagery from textual specifications, generating incident summaries from logs, or producing multimodal training content from fragmented archives. However, two conditions dominate real- world constraints. First, *paired* multimodal data are often scarce or unavailable: text and images may exist in separate repositories, captured by different tools and teams, or governed by distinct policies. Second, data are typically distributed across silos and cannot be aggregated into a single training lake due to privacy, compliance, intellectual property con- straints, and organizational boundaries. Unsupervised and unpaired generation is therefore an essen- tial capability. Image-to-image translation with unpaired data has advanced through cycle-consistency and latent sharing, enabling translation between visual domains without explicit correspondences. Similarly, deep generative modeling has ma- tured through variational and adversarial paradigms, enabling high-quality synthesis. Yet, three practical gaps remain.

- Gap 1: Cross-modal semantic alignment without pairing: Even when unpaired translation works within a modality (e.g., image-to-image), cross-modal synthesis (e.g., text-to-image or audio-to-video) faces an added challenge: semantic structure must be aligned across modalities with fundamentally differ- ent statistical properties. Purely distributional alignment can produce plausible outputs that are semantically inconsistent or difficult to control.
- Gap 2: Trust, integrity, and accountability across silos: Federated learning (FL) enables collaborative training without data sharing by aggregating locally computed updates. Despite progress, FL introduces new attack and failure modes: unre- liable clients, skewed data, poisoning, and accidental miscon- figuration. In enterprise settings, stakeholders require integrity guarantees (the system resists malicious or erroneous updates) and accountability (the system can explain *who* contributed *what*, under what quality standards, without exposing private data).
- Gap 3: Explainability–performance trade-offs: Highly accurate generative models are often opaque. Post-hoc explain- ers can help interpret predictions, but generative pipelines need *controllable* and *auditable* explanations: what factors drove a generated artifact, and which factors can be manipulated to meet policy or business requirements? Moreover, organizations need a *quantifiable* way to decide how much performance they are willing to trade for explainability,

with measurable and optimizable objectives.

1.1. Contributions

This paper addresses these gaps with three contributions:

- ANCHOR-GEN: Latent cross-domain anchors for un- paired multimodal generation. We propose a framework that discovers anchorsa small set of latent factors shared across modalities and domainsand uses them to generate and translate content while maintaining anchor consistency. Anchors serve as a bridge between modali- ties and provide intrinsic explanatory handles.
- Trust metric-based federated learning for integrity and accountability. We introduce a federated framework that assigns each client a dynamic trust score based on multi-signal validation (update consistency, behav- ioral history, robust similarity, and optional cryptographic provenance). Trust scores influence aggregation weights and trigger accountability workflows, providing gover- nance without centralizing data.
- A quantifiable explainability–performance trade-off framework. We define lightweight, non-complex metrics that quantify explainability using anchor stability and concept alignment, and we propose a practical optimiza- tion approach that yields Pareto-efficient operating points suitable for deployment decisions.

1.2. Paper Organization

Section II reviews related work. Section III formalizes the problem. Section IV describes ANCHOR-GEN. Section V introduces trust-metric federated learning for integrity and accountability. Section VI proposes the explainability–performance trade-off framework. Section VII describes ex- periments and results. Section VIII concludes.

2. Related Work

2.1. Deep Generative Modeling

Variational autoencoders (VAEs) provide a probabilistic framework for learning latent representations and generating samples. Adversarial learning, introduced through genera- tive adversarial networks (GANs), improved sample realism by training a generator against a discriminator. Subsequent work improved stability and distribution matching, includ- ing Wasserstein-based formulations. These methods underpin modern generation pipelines but typically do not enforce semantic alignment across modalities without supervision.

2.2. Unpaired and Unsupervised Translation

Unpaired translation methods leverage structural constraints such as cycle-consistency and shared latent spaces to translate between domains without paired examples. Such techniques are powerful within a modality and inspire cross-domain align- ment principles. However, multimodal unpaired translation adds cross-modal semantic mapping challenges, motivating the need for structured shared factors.

2.3. Federated Learning, Privacy, and Robustness

Federated learning formalizes decentralized training where clients compute local updates and a server aggregates them. Secure aggregation protocols enable privacy-preserving sum- mation of updates, while differential privacy mechanisms reduce leakage risks. Robust aggregation methods address adversarial or Byzantine updates through resilient selection or trimming strategies. Enterprise FL additionally demands gov- ernance: trust scoring, accountability, and auditability beyond pure robustness.

2.4. Explainable AI and Trust

Interpretability methods such as LIME and SHAP provide local explanations for model outputs, improving transparency and user trust. However, many explanation techniques are post- hoc and may not translate directly into controllable factors for generation. We aim for intrinsic explainability by construction: anchors represent interpretable latent factors aligned across modalities, enabling both explanation and controllability. Trust is treated as a measurable, dynamic property integrated into training and aggregation.

3. Problem Formulation

3.1. Data Setting

We consider K organizations (clients) participating in fed- erated training. Each client k holds private datasets across multiple modalities and domains. For simplicity, consider two modalities: modality A (e.g., images) and modality B (e.g., text). Each client contains unpaired sets:

- DA = {ai} from domain(s) in modality A
- DB = {bj} from domain(s) in modality B

k
k

There is no assumption of pairing between a_i and b_j, either within a client or across clients. We seek a global model that supports:

- Generation: synthesize samples in each modality from latent factors.
- Cross-modal translation: produce semantically aligned outputs across modalities using shared structure.
- Controllability and explanation: provide human- auditable latent factors that explain generated content.

3.2. System Requirements

The training procedure must meet enterprise deployment constraints:

- Privacy: raw data never leave clients; optional privacy enhancements for updates.
- Integrity: the aggregation resists low-quality, erroneous, or adversarial updates.
- Accountability: contributions are auditable through non- sensitive metadata and trust signals.
- Measurable explainability–performance trade-offs: stakeholders can select operating points.

4. ANCHOR-GEN: Latent Cross-Domain Anchors for Unsupervised Multimodal Generation

4.1. Overview

ANCHOR-GEN introduces *latent cross-domain anchors* as a structured bridge between modalities and domains. Anchors are designed to satisfy three properties:

- P1: Shared semantics. Anchors represent factors that should correspond across modalities (e.g., "sentiment," "category," "style," "severity," "tone").
- P2: Compactness and interpretability. Anchors are small in number and encourage disentangled, stable representations that can be described and audited.
- P3: Learnability from unpaired data. Anchors must be discoverable without paired supervision, using consistency and distributional constraints.

4.2. Model Components

We describe a two-modality instantiation; extension to more modalities follows the same pattern.

- Encoders and decoders: For modality A, an encoder E_A maps an input a to a latent representation. For modality B, E_B maps b to a latent representation. Each modality also has a generator/decoder G_A and G_B that reconstruct samples from latents.
- Anchor extractor: A shared anchor head $H(\cdot)$ maps modality-specific latent representations into an anchor space Z with dimension m, where m is intentionally small (e.g., 8–64). The output $z \in Z$ is the anchor vector.
- Anchor-conditioned generation: Generators are condi-tioned on anchors to produce outputs consistent with anchor semantics. Intuitively, anchors act as "knobs" controlling high- level attributes of generated content.
- Lightweight critics: Instead of complex adversarial objec- tives, ANCHOR-GEN uses lightweight critics or classifiers that encourage (i) within-modality realism and (ii) cross- domain anchor alignment. These critics can be implemented as simple discriminators or domain classifiers.

4.3. Anchor Discovery without Pairing

In the absence of paired data, ANCHOR-GEN uses three complementary training signals.

- Anchor Consistency Under Augmentation: Within each modality, we apply semantic-preserving transformations (e.g., mild cropping for images, paraphrase-like perturbations for text when feasible, or token masking). The anchor representa- tion is encouraged to remain stable under such transformations. This promotes invariance and improves interpretability by forcing anchors to represent persistent factors rather than noise.
- Cross-Domain Anchor Distribution Alignment: Even without pairing, the distribution of anchors extracted from each modality should match at a coarse level for shared semantics. ANCHOR-GEN aligns anchor distributions by encouraging similarity between aggregated anchor statistics (e.g., mean and covariance matching) and by penalizing modality-identifying signals in the anchor space via a simple domain classifier. This discourages anchors from encoding modality-specific artifacts.
- Cycle-Anchor Consistency: For unpaired translation, a sample from modality A is encoded into anchors and then decoded into modality B, encoded again, and required to preserve anchors. The same is applied in the reverse direction. This avoids relying on pixel-perfect or token-perfect cycle reconstruction, focusing instead on semantic cycle consistency through anchors. This is particularly beneficial in cross-modal settings where exact cycle reconstruction is ill-posed.

4.4. Intrinsic Explainability via Anchors

Unlike post-hoc explainers that attempt to interpret black- box behavior, ANCHOR-GEN makes anchors a first-class interface:

- Anchor attribution: for a generated sample, the system reports which anchors were active (e.g., top-r anchor dimensions by magnitude).
- Anchor interventions: users can adjust anchors to con- trol outputs, enabling "what-if" analysis.
- Anchor descriptors: anchors can be labeled using weak signals (e.g., clustering and human naming) without re-training the full model.

This design supports enterprise audit workflows: the model can expose which semantic factors influenced an output without exposing raw training data.

4.5. Training Procedure (Central View)

A conceptual training loop alternates between:

- Within-modality reconstruction to ensure generators re- main grounded.
- Anchor stability constraints under augmentation for each modality.
- Cross-domain anchor alignment constraints to unify se- mantics.
- Cycle-anchor constraints to preserve semantics across translation.

In federated deployment, each client performs local training for a small number of steps and produces an update, which is then aggregated under the trust framework introduced next.

5. Trust Metric-Based Federated Learning for Integrity and Accountability

5.1. Motivation

Standard FL aggregation (e.g., simple averaging) assumes updates are broadly reliable. In practice, organizations may have heterogeneous data, uneven compute, different prepro- cessing, and varying operational maturity. Additionally, mali- cious or compromised clients may attempt poisoning. Robust aggregation improves resilience but does not fully address enterprise governance demands: *why* was an update trusted, *which* signals were used, and *how* can the decision be audited? We propose a trust metric-based federated learning layer that assigns each client a dynamic trust score derived from multiple auditable signals. The trust score influences:

- Aggregation weight,
- Participation frequency,
- Escalation actions (e.g., quarantine, additional validation),
- Accountability logging.

5.2. Trust Signals

Let u_k be the update submitted by client k in a round. We compute a trust score $T_k \in [0, 1]$ as a weighted combination of normalized signals.

- S1: Update consistency. Compare u_k to the cohort using similarity measures (e.g., cosine similarity) computed on a compressed representation of updates. Updates that strongly deviate may be down-weighted.
- S2: Behavioral reliability history. Maintain a client repu- tation profile across rounds: frequency of anomalous updates, stability of contributions, and past validation outcomes. This provides temporal smoothing and discourages sudden harmful shifts.
- S3: Utility validation (privacy-preserving). Evaluate the effect of an update on a small, policy-approved validation set that can be held by the aggregator or distributed as a shared "public" validation resource. When this is not feasible, proxy utility can be estimated via consistency of anchor distributions or reconstruction quality on locally held validation splits reported as aggregates.
- S4: Robustness checks. Apply robust aggregation rules as a first filter (e.g., Krum-like selection or Bulyan-style refinement) and treat pass/fail or rank as a trust signal rather than as the only decision mechanism.
- S5: Provenance and audit metadata. Clients optionally attach non-sensitive metadata: model version, preprocessing hash, training step count, and secure attestation tokens when available. These do not reveal data but improve accountability.

5.3. Aggregation with Trust Weighting

Trust scores determine aggregation weights. The key design goal is *bounded influence*: even highly trusted clients should not dominate, while low-trust clients should have sharply reduced impact.

A practical rule is:

- Clip each update norm to reduce scale manipulation,
- Assign a weight based on trust, capped to a maximum ratio, aggregate using weighted averaging after robust filtering. This creates a hybrid approach: robust filtering protects against extreme outliers, and trust weighting improves governance and resilience under milder but persistent issues.

5.4. Integrity Controls

- Secure aggregation: To prevent the server from inspecting individual updates, secure aggregation can be used to compute sums without revealing per-client contributions. In such cases, some trust signals must rely on client-reported metrics or on limited, privacy-safe sketches. The framework supports both modes: full visibility for controlled consortia, or privacy- preserving mode with reduced observability.
- Differential privacy (optional): Clients may add noise to updates and enforce clipping to reduce information leakage.

Trust scoring can be adjusted to account for privacy-induced variance, avoiding penalizing privacy-compliant clients.

- Byzantine resilience: When adversarial risk is high, ro- bust methods (e.g., Krum/Bulyan families) are integrated.

Importantly, the trust framework does not replace robustness; it complements it with accountability signals and enterprise controls.

5.5. Accountability Layer

Accountability requires that model governance teams can reconstruct:

- Which clients participated per round,
- What trust signals were used,
- How trust scores were derived,
- What aggregation decision was made.

We propose maintaining an *audit log* containing per-round metadata and trust summaries. To avoid storing sensitive information, the log includes hashed identifiers, trust scores, signal summaries, and policy decisions (e.g., "quarantined" or "down-weighted"). This enables retrospective audits and continuous improvement.

6. Quantifying and Optimizing the Explainability–Performance Trade-off

6.1. Rationale

Organizations frequently face the question: *How much per- formance should we trade for explainability?* In practice, the trade-off must be quantified in operational terms using metrics that are:

- Lightweight (computable in federated settings),
- Stable (robust to randomness),
- Actionable (improvable via training controls).

ANCHOR-GEN provides a natural interface for quantification: anchors can be assessed for stability and semantic alignment, while generation quality can be evaluated using standard metrics.

6.2. Explainability Metrics (Anchor-Based)

We propose three explainability metrics that avoid complex mathematics.

- E1: Anchor stability. Measure how consistent anchors re- main under semantic-preserving augmentations. High stability implies anchors represent meaningful factors rather than noise. E2: Cross-modal anchor agreement. For translated sam- ples, compare anchor vectors before and after translation cy- cles. Higher agreement indicates semantic preservation across modalities.
- E3: Anchor sparsity and compactness. Encourage expla- nations that rely on fewer anchors per sample. A compact explanation is easier for humans and auditors to interpret. This can be measured as the fraction of anchor dimensions exceeding a small threshold.

These metrics can be computed locally and aggregated as summary statistics in FL.

6.3. Performance Metrics

Performance is measured using task-appropriate genera- tion/translation quality metrics. In unpaired translation and generation, widely used metrics include:

- Distribution similarity scores such as FID (for image domains).
- Reconstruction fidelity where applicable.
- Simple downstream utility: e.g., classifier accuracy on generated samples or retrieval accuracy when embeddings are available.

For multimodal settings involving text, we emphasize quali- tative and lightweight quantitative proxies that do not require complex language metrics.

6.4. Optimization Strategy

We treat training as a multi-objective problem: maximize performance while maximizing explainability. In practice, en- terprise teams benefit from a small set of operating points rather than an abstract Pareto frontier. We propose:

- Scalarization with policy weights: set a policy weight λ that indicates the relative importance of explainability,
- Anchor budget scheduling: begin with stronger explain- ability regularization (to stabilize anchors early), then relax to recover performance,
- Trust-aware tuning: when trust is low (high risk), favor explainability to improve auditability; when trust is high, allow more performance-focused training.

6.5. Deployment Decision Support

We recommend reporting a simple dashboard per model version:

- Performance score (e.g., FID or proxy utility),
- Explainability score (aggregate of E1–E3),
- Trust health score (aggregated T_k distribution across clients),
- An operating point label (e.g., "High Explainability", "Balanced", "High Performance").

This turns model selection into a governance-aligned process.

7. Experimental Setup

7.1. Goals

Experiments evaluate:

- G1: semantic consistency in unpaired cross-domain and cross-modal generation,
- G2: explainability improvements from anchors,
- G3: resilience and accountability benefits from trust- metric FL,
- G4: measurable explainability–performance trade-offs.

7.2. Datasets and Tasks

We use standard, conceptually clear benchmarks suitable for unpaired settings. To keep tables narrow, we summarize datasets and tasks succinctly.

The "synthetic captions" setting uses text pools derived from attribute templates to emulate enterprise text silos where text

Table 1: Datasets and Unpaired Tasks (Summary)

Dataset	Task (Unpaired)
MNIST, SVHN	Cross-domain image translation (digits)
CelebA subsets	Attribute/style translation across partitions
Synthetic captions	Cross-modal anchor alignment with un-paired text/image pools

Exists without direct pairing to images. This allows evaluation of cross-modal anchor alignment without requiring paired caption datasets.

7.3. Baselines

We compare against representative unpaired translation and latent-sharing paradigms:

- Unpaired translation with cycle-consistency (within- modality),
- Shared-latent unpaired translation frameworks,
- Federated averaging (no trust scoring),
- Robust aggregation variants (robust filtering without ac- countability scoring).

7.4. Federated Simulation

We simulate K clients by partitioning datasets into non- identical splits. To emulate enterprise heterogeneity:

Clients have different label/attribute distributions, some clients have smaller datasets,some clients apply slight preprocessing shifts. We introduce two fault modes:

- Noisy client: poor preprocessing causing low-quality updates,
- Adversarial client: poisoned updates that attempt to destabilize anchors or degrade translation.

7.5. Evaluation Metrics

Performance:

- FID for image-domain tasks where applicable,
- Proxy utility via a simple classifier trained on real data and evaluated on generated samples,
- Cycle-anchor agreement for translation quality.

Explainability:

- Anchor stability under augmentations (E1),
- Cross-modal anchor agreement (E2),
- Anchor sparsity/compactness (E3).

Trust and integrity:

- Degradation under faults (relative performance drop).

- Fraction of rounds where faulty updates are down- weighted.
- Audit completeness (availability of per-round trust sum- maries).

Table 2: Explainability–Performance Trade-off (Illustrative Summary)

Setting	Perf.	Explain.
High Perf.	Best	Low–Med
Balanced	Near-best	High
High Explain.	Med	Best

8. Results and Discussion

8.1. Unpaired Translation and Anchor Consistency

ANCHOR-GEN improves semantic consistency by shifting cycle constraints from raw reconstruction to anchor preserva- tion. In practice, this reduces failure cases where translations look plausible but do not preserve intended factors. Across digit translation tasks, anchors become strongly associated with digit identity and style factors. In attribute translation settings, anchors correlate with interpretable dimensions (e.g., "smile intensity" or "hair tone") even when clients see differ- ent attribute distributions.

8.2. Explainability Measurements

Anchor stability increases consistently compared to base- lines that do not explicitly enforce anchor invariance. This matters operationally: if explanations vary wildly under trivial input perturbations, they are difficult to trust. Anchor sparsity improves when regularization is enabled, yielding explanations that rely on a small subset of anchors.

8.3. Trade-off Curves without Complex Machinery

A key observation is that modest explainability regulariza- tion yields large gains in anchor stability with limited perfor- mance loss. As regularization increases further, explainability continues to improve but performance begins to degrade. This motivates selecting balanced operating points for enterprise deployment. Table II summarizes the typical behavior observed: a balanced setting often retains near-best performance while substantially improving explainability. This is valuable for governance contexts where explanations are mandatory.

8.4. Federated Trust: Integrity and Accountability

Under noisy or adversarial clients, plain federated averag- ing suffers measurable degradation: anchors become unstable and cross-domain consistency deteriorates. Robust aggregation alone improves resilience but may be opaque to auditors. The proposed trust layer improves both resilience and governance:

- Faulty updates are down-weighted more consistently due to multi-signal trust scoring.
- Clients with persistent issues accumulate low trust, re- ducing long-term impact.
- Audit logs capture per-round decisions, enabling post- incident analysis.

In practice, the trust framework helps reconcile two enterprise needs that are often in tension: privacy-preserving collabora- tion and accountable governance.

8.5. Practical Considerations

- Anchor dimension: Very small anchor spaces may underfit semantics; very large anchor spaces reduce interpretability. A moderate anchor dimension offers a pragmatic balance.
- Client heterogeneity: Non-IID data can cause anchor drift. Trust-aware aggregation mitigates but does not eliminate this; anchor regularization and periodic global calibration help.
- Privacy vs observability: Secure aggregation reduces the server's ability to compute some trust signals directly. In privacy-preserving mode, we recommend relying more on client-reported aggregate metrics and robust filtering, while keeping accountability via metadata and round summaries.

9. Conclusion

This paper introduced ANCHOR-GEN, an unsupervised multimodal generation framework based on latent cross- domain anchors that align semantics across modalities without requiring paired data. Anchors provide intrinsic explainability and controllability by design, enabling audit-friendly genera- tion pipelines. To support enterprise deployment across silos, we proposed a trust metric-based federated learning framework that enforces integrity through robust weighting and provides accountability via auditable trust summaries and provenance- aware governance. Finally, we presented a practical method to quantify and optimize the explainability–performance trade- off using lightweight anchor-based metrics, enabling selection of governance-aligned operating points.

Future work includes extending anchors to richer hierar- chical structures, improving privacy-preserving trust signals

under secure aggregation constraints, and evaluating anchor- based governance in real enterprise workflows with human auditors.

Acknowledgment

The author thanks the broader research community for foundational work in federated learning, privacy, robust aggregation, deep generative modeling, and interpretable machine learning.

References

[1] M. Gutmann and A. Hyva¨rinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. AIS- TATS*, 2010.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their composi- tionality," in *Proc. NeurIPS*, 2013.

[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NeurIPS*, 2016.

[6] V. Dumoulin *et al.*, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[7] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM CCS*, 2016. :contentReference[oaicite:0]index=0

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, 2016. :contentReference[oaicite:1]index=1

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adver- sarial networks," in *Proc. ICML*, 2017.

[10] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agu¨era y Arcas, "Communication-efficient learning of deep net- works from decentralized data," in *Proc. AISTATS*, 2017. :contentRe- ference[oaicite:2]index=2

[11] K. Bonawitz *et al.*, "Practical secure aggregation for privacy- preserving machine learning," in *Proc. ACM CCS*, 2017. :contentRe- ference[oaicite:3]index=3

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017.

[13] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. NeurIPS*, 2017.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to inter- preting model predictions," in *Proc. NeurIPS*, 2017. :contentRefer- ence[oaicite:4]index=4

[15] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc.NeurIPS*, 2017. :contentReference[oaicite:5]index=5

[16] X. Huang *et al.*, "Multimodal unsupervised image-to-image translation," in *Proc. ECCV*, 2018.

[17] E. M. El Mhamdi, R. Guerraoui, and S. Rouault, "The hidden vul- nerability of distributed learning in Byzantium," in *Proc. ICML*, 2018. :contentReference[oaicite:6]index=6

[18] P. Kairouz *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019. :contentRefer- ence[oaicite:7]index=7