BoMM: Multi-Modality Large-Small Model Bidirectional Collaboration

Anonymous authors
Paper under double-blind review

Abstract

Different from existing single-modality large-small model collaborations, multi-modality large-small model collaboration is an under-explored paradigm where cloud-side multi-modality large model (MM-LM) collaborates with parties' small models (SMs) to achieve bidirectional domain-specific performance improvements. Nevertheless, this paradigm faces two key challenges. First, MM-LM inherently relies on abundant modality-aligned samples for training, but geographical and device diversity across parties inevitably lead to different collected samples and modalities. These differences significantly reduce overlapping sample entities across parties' multi-modality datasets, creating **modality alignment scarcity** challenge. Second, collected device failure and human annotation costs further lead to different modality missing problems in each party's dataset. Existing modality completion methods typically require enough modality-completed training samples to ensure generation quality, creating a **modality completeness gap** challenge. To address these challenges, we propose a **m**ulti-**m**odality large-small model **b**idirectional c**o**llaboration framework, named **BoMM**, which consists of two key components. Specifically, *global prototype-guided alignment* strategy identifies potentially aligned samples through similarity distribution comparisons between unaligned data and established global prototypes, enabling knowledge transfer from SMs to MM-LM. With established prototypes, *preference-driven modality adaptive completion* method integrates direct preference optimization into generator training with real-time scheduling to dynamically complete missing modalities, enabling knowledge transfer from MM-LM to SMs. Theoretical analysis confirms BoMM's $O(1/\sqrt{T})$ convergence rate. Across three multi-modality scenarios, it outperforms state-of-the-art methods by up to 6.64% on two well-known datasets. Our code is available at https://anonymous.4open.science/r/MultiLM-5D65.

1 Introduction

In large language model (LLM) and small models (SMs) collaborative scenarios, cloud-side LLM leverages parties' domain data and SMs' expertise to address domain-specific weaknesses, while SMs of parties can also benefit from LLM' superior reasoning and generation capabilities Wang et al. (2024a); Liu et al. (2024). This collaborative relationship extends to federated learning scenarios Guo et al. (2024) when privacy becomes essential Zhang et al. (2025), enabling mutual enhancement while keeping all raw data local to satisfy privacy requirements.

Existing large-small model collaborations utilize three primary approaches: (i) distillation-based methods Zhou et al. (2021); Liu et al. (2021); Pham et al. (2021) leverage knowledge distillation to transfer knowledge between domain-specific SMs and LLM in both directions; (ii) generation-based methods Zhang et al. (2021); Li & Jin (2022); Nasser et al. (2024) leverage synthetic datasets to transfer knowledge from SMs to LLMs or use LLMs' generative capabilities to create task-specific datasets; and (iii) parameter-based methods Fan et al. (2024); Cheng et al. (2021); Yu et al. (2023) employ parameter-efficient fine-tuning (PEFT) techniques Han et al. (2024) to transmit knowledge bidirectionally or selectively transfer parametric knowledge from LLMs to SMs.

Despite significant achievements, these methods focus on single-modality large-small model collaboration, neglecting multi-modality scenarios. In reality, with multi-modality data becoming more

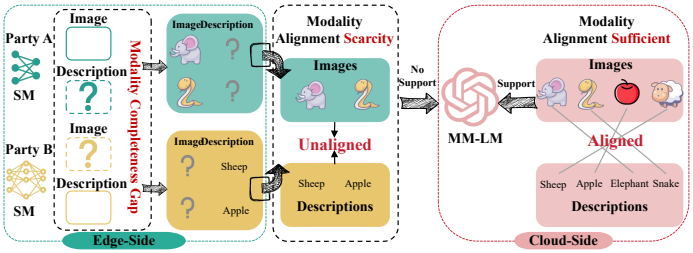

Figure 1: Multi-modality large-small model collaboration face to key challenges: modality completeness gaps from parties' modality missing, and modality alignment scarcity from parties' modality alignment problems for the same sample, hindering MM-LM training that relies on abundant modality alignments.

prevalent, multi-modality large model (MM-LM) are becoming an increasingly important research direction Tong et al. (2025). Since MM-LM often exhibits suboptimal performance on domain-specific tasks Ye et al. (2024), they require access to multi-modality domain data from parties. This necessitates collaboration between multi-modality large and small models. Although a recent study Chen et al. (2025) aligns representations between single-modality MM-LM and multi-modality SMs in medicine, it relies heavily on two idealistic assumptions: modality alignment and modality completeness for each sample. These rarely hold in real-world distributed settings. Additionally, this work also overlooks the need for mutual performance improvements in multi-modality collaborations, where the limitation of mutual performance improvement is also observed in single-modality collaborations.

Therefore, we explore an under-explored paradigm: multi-modality large-small model bidirectional collaboration, as shown in Fig. 1. Different from existing large-small model collaborations, due to geographical and device diversity across parties, this paradigm faces a **modality alignment scarcity** challenge where parties possess multi-modal data but lack aligned samples across modalities. These aligned samples are essential for training MM-LM's alignment capability Wu et al. (2023); Song et al. (2023). Moreover, we identify an additional critical challenge: **modality completeness gap**, which occurs when parties have incomplete or missing modalities due to device failures or annotation expenses. Existing completion approaches Wu et al. (2024a) still struggle to generate high-quality modalities with severe or even complete modality missing in training datasets.

To address these challenges, we propose a *m*ulti-*m*odality large-small model **b**idirectional c**o**llaboration framework, denoted as **BoMM**, where a *global prototype-guided alignment* method identifies potentially aligned samples by comparing similarity distributions to established global prototypes, enabling knowledge transfer from SMs to MM-LM. Building on these prototypes, *preference-driven modality adaptive completion* method optimizes the global generator to dynamically fill local missing modalities with a real-time sample scheduler, facilitating effective knowledge transfer from MM-LM to SMs. The main contributions are summarized as follows:

(a) We present an under-explored paradigm: **m**ulti-**m**odality large-small model **b**idirectional c**o**llaboration, and propose BoMM, where cloud-side MM-LM enhances domain performance by SMs, while parties' SMs overcome modality missing using MM-LM's generation capability.

(b) We mitigate modality alignment scarcity challenge through a global prototype-guided alignment method, and address modality completeness gap challenge via a preference-driven modality adaptive completion method, where a real-time sample scheduler is designed to filter generated modalities, enabling effective bidirectional knowledge transfer between models.

(c) BoMM can effectively handle different modality alignment situations, while supporting various modality missing rates of parties. Our theoretical analysis guarantees convergence at $O(1/\sqrt{T})$ rate, ensuring robust performance in multi-modality large-small model collaborations.

(d) Extensive experiments on two real-world multi-modal datasets across three scenarios clearly demonstrate that BoMM achieves substantial improvements over state-of-the-art methods under varying rates of aligned samples and modality missing.

2 Related Work

2.1 Large-Small Model Collaboration

large-small model collaboration leverages distributed private domain data to address public data scarcity and domain-specific poor performance challenges in large language models (LLMs). Con-

currently, the extensive knowledge within LLMs can improve complex reasoning capabilities and provide comprehensive background information for domain-specific small models (SMs). Three primary methods facilitate knowledge transfer between SMs and LLMs Liu et al. (2025): distillation-based Zhou et al. (2021); Liu et al. (2021); Pham et al. (2021), generation-based Zhang et al. (2021); Li & Jin (2022); Nasser et al. (2024), and parameter-based transfer methods Fan et al. (2024); Cheng et al. (2021); Yu et al. (2023). These approaches primarily concentrate on unidirectional knowledge transfer between large-small models, with few works Fan et al. (2024); Cheng et al. (2021) exploring bidirectional knowledge exchange. More importantly, existing works fail to cope with collaborative scenarios for a multi-modality large model (MM-LM) and SMs, where parties may possess heterogeneous data modalities, diverse model architectures, and face varying degrees of modality incompleteness in their local datasets.

2.2 Large multi-modality model

The rapid advancement of multi-modality large model has significantly demonstrated their remarkable ability to tackle various multi-modality challenges in real-world applications, including personal AI assistants Wang et al. (2024b), education Lee et al. (2025), medicine Wang et al. (2024c), autonomous driving Cui et al. (2024), and others. However, these works depend on training data where multiple modalities (such as images and text) are paired for each sample, a condition rarely met in distributed environments. In these distributed environments, each party typically holds only specific modalities for their local samples, while complementary modalities for these exact samples are often entirely absent from other parties' collections. This modality misalignment generally emerges from regional differences, domain specializations, and other organizational factors, making complete multi-modality pairing exceedingly difficult to achieve. Leveraging such unaligned multi-modality data for MM-LM training can lead to weakened multi-modality understanding, increased hallucinations, and unreliable correlations between different modalities, ultimately limiting the model's practical utility in complex real-world scenarios Song et al. (2023).

3 Preliminary

3.1 Multi-modality large-small model collaboration

In our multi-modality large-small model collaboration scenarios with K parties and modality set M, each party $k \in \{1, 2, ..., K\}$ has dataset D_k containing N_k samples and $M_k \subseteq M$ modalities. Parties may have identical or different modalities due to modality missing, with different sample alignment ratios. Within dataset D_k, each party has labeled samples $(x_i^l, y_i^l) \in D_k^l$ and unlabeled samples $(x_i^{ul}, y_i^{ul}) \in D_k^{ul}$. These data include aligned samples $(x_i^a, y_i^a) \in D_k^a$ (same sample ID across K parties) and unaligned samples $(x_i^{ua}, y_i^{ua}) \in D_k^{ua}$ (sample ID found only in party k).

3.2 Optimization Objective

Resource limitations lead parties to deploy lightweight small models (SMs) locally while collaboratively training a powerful domain-specialized multi-modality large model (MM-LM) on the cloud. Our framework establishes bi-directional knowledge transfer, simultaneously incorporating domain-specific expertise from SMs into MM-LM while enabling MM-LM to enhance SMs' performance. Formally, we denote MM-LM with parameters ψ as f_ψ and SM of party k with parameters θ_k as f_{θ_k}. Our bi-directional optimization objective includes:

Global MM-LM: The cloud-side multi-modality large model f_ψ is optimized using parties' multi-modality data for both domain-specific tasks and modality completion:

$$\min_{\psi} \mathcal{L}_{\text{global}}(\psi; \{D_k\}_{k=1}^K, \{\theta_k\}_{k=1}^K), \tag{1}$$

where $\psi = [\psi_1, \psi_2]$, with ψ_1 and ψ_2 representing classifier and generator parameters, respectively.

Local SMs: Each party's small model f_{θ_k} is optimized using its local dataset D_k for domain-specific tasks, improved by enhanced domain data from MM-LM:

$$\min_{\theta_k} \mathcal{L}_{\text{local}}(\theta_k; D_k, \psi), \tag{2}$$

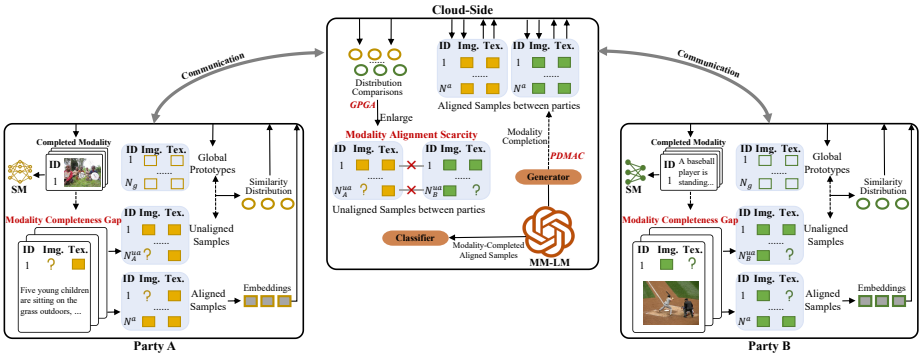

Figure 2: The overview of BoMM framework. The global prototype-guided alignment (GPGA) method identifies essential modality-aligned samples by analyzing similarity distributions between unaligned data and global prototypes, enhancing MM-LM's domain performance. Meanwhile, parties leverage a preference-driven modality adaptive completion (PDMAC) method to synthesize missing modality using MM-LM's generation ability, improving local domain performance.

where $\mathcal{L}_{local}$ denotes the party-specific optimization loss for each local model, and $\mathcal{L}_{global}$ represents the optimization loss for global MM-LM in the cloud.

4 Our Methodology

In this section, we propose BoMM, a multi-modality large-small model bidirectional collaboration framework including a preference-driven modality adaptive completion method with a real-time sample scheduler and a global prototype-guided alignment strategy. The overall architecture is illustrated in Fig. 2.

4.1 Preference-Driven Modality Adaptive Completion

The modality incompleteness of each party degrades local model performance on private domain tasks while simultaneously limiting knowledge contribution to cloud-side MM-LM training. Resource constraints prevent parties from training domain-specific MM-LM locally for modality completion. While general MM-LM is accessible via APIs, they generate generic outputs that fail to synthesize missing modalities for specialized domains accurately.

4.1.1 Modality Adaptive Completion

Inspired by direct preference optimization (DPO) Rafailov et al. (2023), which guides models to generate human-preferred outputs, we design a preference-driven modality adaptive completion (PDMAC), enabling MM-LM to generate domain-specific modality information tailored to local private data. Unlike existing DPO approaches Wu et al. (2024b); Karthik et al. (2024) for modality completion that require extensive modality-completed training datasets, our method effectively handles different modality incompleteness or entire modality missing. By leveraging MM-LM for modality completion, we substantially improve model performance on domain-specific tasks.

To be specific, for i-th sample with existing modality information $x_{i,k}^{\{v,t\}}$, we create a preference pair (g_i^p, g_i^q). g_i^p is the preferred output, which is obtained from the closest global prototype o_i to the i-th sample, while g_i^q is the non-preferred output, which is the missing modality generated by the MM-LM's generator $\mathcal{G}$. The optimization objective for $\mathcal{G}$ is:

$$\mathcal{L}_{\text{prefer}} = -\mathbb{E}_{(x_{i,k}^{\{v,t\}}, g_i^p, g_i^q)\sim D_k}\left[\log\sigma(\beta(R(x_{i,k}^{\{v,t\}}, g_i^p)) - R(x_{i,k}^{\{v,t\}}, g_i^q))\right], \tag{3}$$

where $\{v,t\}$ indicates available image (v), text (t), or both modalities. σ is the logistic function, and β scales the reward function $R(\cdot)$. Since parties may have different missing modalities, $R(\cdot)$ is adapted based on the generated modality type, following Jiang et al. (2024) for text and Karthik et al. (2024) for images.

Additionally, to prevent generated modality from lacking personalization and becoming too similar to its global prototype, we introduce an additional loss term:

$$\mathcal{L}_{\text{special}} = \|g_i^q - g_i^p\|^2 (1 - \text{sim}(x_{i,k}^{\{v,t\}}, o_i^{\{v,t\}})), \tag{4}$$

where o_i and $x_{i,k}^{v,t}$ share modalities i-th sample already has, while g_i^q and g_i^p represent the target modality to generate. Our method needs just one complete-modality sample per class as a prototype, which is practical to obtain. If unavailable initially, existing generation methods Kwon et al. (2024); Laina et al. (2019) still can initialize these prototypes. Once established, samples grouped around these prototypes progressively receive high-quality completion results during training. For samples already obtaining the target modality, these real data serves directly as preferred output g_i^p. The generator $\mathcal{G}$ is updated by $\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\text{prefer}} + \gamma \mathcal{L}_{\text{special}}$, where γ is a balancing hyperparameter. The MM-LM classification task loss term $\mathcal{L}_{task}$ is calculated by the standard cross-entropy loss Zhang & Sabuncu (2018) with completed modality data.

4.1.2 Dynamic sample scheduler

Parties use the cloud-side MM-LM to generate missing modalities and expand local datasets. To handle potentially noisy generated data, we implement a scheduler that dynamic select high-reward samples in each batch based on sample reward. Low-quality samples are kept in the training set, with their current generations continuing to serve as non-preferred outputs for further optimization.

Reward definition. Since task loss can be applied to measure sample value Huang et al. (2025), we design a reward function that dynamically evaluates generator improvement after each batch update. Our reward function $r(\cdot)$ combines sample-level and batch-level assessments, with batch-level reward measuring the reduction in average generated loss before and after training. In practice, we use batch-average generated loss as an approximation of the expected generated loss across the entire data distribution. For the i-th sample in batch b, the reward is:

$$r(a_k^{i,b}, \theta_{\mathcal{C}_k}^{b-1}, \theta_{\mathcal{C}_k}^{b}) = \underbrace{-\log \mathcal{P}_{\theta_{\mathcal{C}_k}^{b-1}}(y_k^i | \bar{x}_k^{i,b}) + \delta \mathcal{H}}_{\text{sample-level}} + \underbrace{\frac{\overbrace{\sum_{i=1}^{N_b} e^{\mathcal{L}_{\mathcal{G}}(a_k^{i,b}; \theta_{\mathcal{C}_k}^{b-1})}}^{A} - \overbrace{\sum_{i=1}^{N_b} e^{\mathcal{L}_{\mathcal{G}}(a_k^{i,b+1}; \theta_{\mathcal{C}_k}^{b})}}^{Q}}{\max\left(\sum_{i=1}^{N_b} e^{\mathcal{L}_{\mathcal{G}}(a_k^{i,b}; \theta_{\mathcal{C}_k}^{b-1})}, \sum_{i=1}^{N_b} e^{\mathcal{L}_{\mathcal{G}}(a_k^{i,b+1}; \theta_{\mathcal{C}_k}^{b})} \right)}}_{\text{batch-level}}, \tag{5}$$

where $a_k^{i,b} = (\bar{x}_k^{i,b}, y_k^i)$. $\theta_{\mathcal{C}_k}^{b-1}$ represents classifier parameters in batch b, and $\mathcal{P}_{\theta_{\mathcal{C}_k}^{b-1}}$ is the model's output probability distribution given input $a_k^{i,b}$. N_b is the batch size. Term A evaluates $\theta_{\mathcal{C}_g}^{b-1}$ on batch $b-1$, while term Q evaluates $\theta_{\mathcal{C}_g}^{b+1}$ on batch $b+1$. $\bar{x}_k^{i,b}$ combines intermediate representations across all layers to utilize the model's full representational capacity.

Scheduler design. The rewards of generated samples serve as supervisory signals for training the dynamic scheduler $\mathcal{R}$. This scheduling process addresses the exploration-exploitation trade-off Berger-Tal et al. (2014), requiring a balance between discovering new sample spaces and utilizing known high-quality examples. We implement two networks after an encoder layer: R^p for exploitation and R^q for exploration.

Specifically, the exploit network $\mathcal{R}^p$ predicts sample rewards by mapping inputs to observed rewards, while the explore network $\mathcal{R}^q$ estimates uncertainty and adds an exploration bonus. This design balances exploitation and exploration during sample selection, following UCB Wen et al. (2015) and Thompson Sampling principles Zhang et al. (2020). For modality incomplete data $\bar{x}_k^{i,b} \in D_k^{\text{inc}}$ in b-th batch, $\mathcal{R}^p$ eedforward neural network with residual connections, denoted by $\mathcal{R}^p(\bar{x}_k^{i,b}; \theta_p^b)$. After obtaining the observed reward $r(a_k^{i,b}, \theta_p^{b-1}, \theta_p^b)$, the parameters θ_p^b are updated by:

$$\mathcal{L}^p(D_{k,b}^{\text{inc}}, \theta_p^b) = \frac{1}{2N_b} \sum_{i=1}^{N_b} \left[\mathcal{R}^p(\bar{x}_k^{i,b}; \theta_p^b) - r(a_k^{i,b}, \theta_p^{b-1}, \theta_p^b) \right]^2. \tag{6}$$

The explore network $\mathcal{R}^q$ takes input $h_{i,b,k}^p$ created by concatenating intermediate hidden states of $\mathcal{R}^p(\bar{x}_k^{i,b}; \theta_p^{b-1})$ with the last dimension. This allows $\mathcal{R}^q$ to consider the exploit network's internal

states. Like $\mathcal{R}^p$, $\mathcal{R}^q$ is a feedforward neural network with residual connections. Its parameters θq^b are updated by:

$$\mathcal{L}^q(D^{\text{inc}}_{k,b}, \theta^b_q) = \frac{1}{2N_b}\sum_{i=1}^{N_b}\left[\mathcal{R}^q(h^p_{i,b,k};\theta^b_q) - \left(r(a^{i,b}_k, \theta^{b-1}_p, \theta^b_p) - \mathcal{R}^p(\bar{x}^{i,b}_k;\theta^b_p)\right)\right]^2. \tag{7}$$

The final reward estimation combines both networks:

$$\hat{r}(\bar{x}^{i,b}_k;\theta^b_p,\theta^b_q) = \mathcal{R}^p(\bar{x}^{i,b}_k;\theta^b_p) + \lambda\mathcal{R}^q(h^p_{i,b,k};\theta^b_q), \tag{8}$$

where λ controls exploration strength. Samples with predicted rewards $\hat{r} > \mu$ are selected as high-value samples for the current batch, which μ is a hyperparameter.

4.2 Global Prototype-Guided Alignment

Finding aligned data across parties is difficult due to geographic and domain differences. This creates a significant challenge for MM-LM training, which requires aligned multi-modality data. Current alignment techniques Feng (2022); Feng et al. (2022) typically need some pre-existing aligned data and can't handle scenarios with variable alignment quantities, especially with zero-aligned samples. While Guo et al. (2025) addresses zero-aligned scenarios, it assumes uniform modality distribution across parties and complete modalities within samples, failing in multi-modality collaboration settings. To address these limitations, we propose a global prototype-guided alignment (GPGA) strategy that works with any number of pre-existing aligned data, including zero. We use optimal transport (OT) distance to derive relationship vectors between unaligned samples and global prototypes, then identify new aligned data by analyzing these relationship vectors.

Specifically, even with zero pre-existing aligned data, we start by using local labeled data D^l_k to create class prototypes $\{p^c_k\}^C_{c=1}$ through weighted averaging Bien & Tibshirani (2011). We pair same-class prototypes across parties as initial aligned data to seed the global prototype set $\mathcal{O} = \{o^m | m = 1, ..., N_{\text{global}}\}$. Using optimal transport Peyré et al. (2019), we calculate transport cost $M^k_{im} = 1 - \cos(x^{ua}_i, o^m)$ between unaligned samples and global prototypes. For each unaligned sample, we create a cost vector $\mathbf{M}^k_i$ representing its relationship to all global prototypes. To find corresponding samples across parties, we compute:

$$j^\star = \text{argmax}_{j\in N^{ua}_w}\text{sim}(x^{ua,i}_k, x^{ua,j}_w) = \text{argmax}_{j\in N^{ua}_w}\frac{\mathbf{M}^k_i \cdot \mathbf{M}^w_j}{\|\mathbf{M}^k_i\|_2 \cdot \|\mathbf{M}^w_j\|_2}, \tag{9}$$

where N^{ua}_w is the size of the unaligned dataset D^{ua}_w in party w. This identifies cross-party alignment by comparing relationship patterns to shared prototypes rather than directly comparing incompatible modalities. The newly identified aligned sample pairs join the global prototype set until reaching N_{global}, after which these prototypes guide alignment of remaining samples. These enlarged aligned samples are utilized to improve the domain performance of local SM $\mathcal{L}^k_{task}$ by standard cross-entropy loss Zhang & Sabuncu (2018). These components support each other cyclically: alignment relies on completion for consistent representations, while completion depends on the global prototypes from alignment.

4.3 Complexity and Convergence Analysis

We analyze the convergence of our proposed BoMM. Let $\Psi := \{\psi\}$ and $\Theta := \{\theta_1, \theta_2, \cdots, \theta_{K-1}, \theta_K\}$. The optimization problems in Eq. 1 and Eq. 2 thus form a bilevel optimization problem, which can be formulated as follows.

$$\begin{aligned}\min_{\Psi} F(\Psi) := \mathcal{L}_{\text{global}}(\Psi, \Theta^*(\Psi)) = \mathbb{E}_\xi[L_{\text{global}}(\Psi, \Theta^*(\Psi);\xi)] = \frac{1}{n}\sum_{i=1}^{n} L_{\text{global}}(\Psi, \Theta^*(\Psi);\xi_i),\\ \text{s.t.}\quad \Theta^*(\Psi) = \arg\min_{\Theta}\mathcal{L}_{\text{local}}(\Psi,\Theta) = \mathbb{E}_\zeta[L_{\text{local}}(\Psi,\Theta;\zeta)] = \frac{1}{n}\sum_{i=1}^{n} L_{\text{local}}(\Psi,\Theta;\zeta_i),\end{aligned} \tag{10}$$

where $\mathcal{L}_{\text{global}}$ contains $\mathcal{L}_{\mathcal{G}}$ and $\mathcal{L}_{\text{task}}$. $\mathcal{L}_{\text{local}}$ consists only of $\mathcal{L}^k_{\text{task}}$. L_{global} and L_{local} are loss functions per sample for cloud and parties, where $L_{\text{local}}(\Psi,\Theta;\zeta) := l_{\text{local}}(\Psi,\Theta;\zeta) + \frac{\varphi}{2}\|\Theta\|^2_F$, with ζ representing a sample from K parties. To analyze convergence, we introduce the following definition.

Definition 1. A point $\hat{\Psi}$ is called an ϵ-*accurate stationary point* for the objective function $F(\Psi)$ if $\mathbb{E}\|\nabla F(\hat{\Psi})\|^2 \leq \epsilon$.

Proof Sketch. To explore essential insights, we first bound the tracking error $\|\Theta_t^{j-1} - \Theta^*(\Psi_t^0)\|$ between local parameters and optimal parameters at the j-th epoch. Then, using virtual updates technology (Yang et al., 2022), we establish an upper bound for the gradient approximation error $\left\| \frac{\partial \mathcal{L}_{\text{local}}(\Psi_t^j, \Theta_t^\tau)}{\partial \Psi_t^j} - \nabla F(\Psi_t^j) \right\|$.

Theorem 1. Under Assumptions 1-4 (detailed in Appendix B.1), define $\alpha := -L + \varphi$, choose step size η to be $\frac{2}{L_2+\alpha}$, $N_b = \mathcal{O}(\frac{1}{\sqrt{\epsilon}})$, $\tau = \mathcal{O}(\log \frac{1}{\epsilon})$, $\eta' < \frac{1}{2L_0}$ and suppose $\alpha < L_2$, we have:

$$\frac{1}{\tau' T} \sum_{t=0}^{T-1} \sum_{j=0}^{\tau'-1} \|\nabla F(\Psi_t^j)\|^2 \leq \frac{F(\Psi_0^0) - \inf_\Theta F(\Psi)}{\tau' T(\frac{\eta'}{2} - L_0 \eta'^2)} + (\eta' + 2\eta'^2 L_0) L_2^2 \Delta^2 \frac{\tau' - 1}{\tau'} + \mathcal{O}(\epsilon), \quad (11)$$

where $\mathcal{O}(\cdot)$ indicates growth rate proportional to or slower than ϵ. With $L_0 := L_2 + \frac{2L_2^2 + L_1^2 L_3}{\alpha} + \frac{L_1 L_2 L_3 + L_1 L_2 L_4 + L_2^3}{\alpha^2} + \frac{L_1 L_2^2 L_4}{\alpha^3}$, setting $\tau' = 1$, yields a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$.

5 Experiments

5.1 Experimental Setup

Settings: We evaluate our proposed BoMM framework on two widely-used multi-modal datasets: MS COCO Lin et al. (2014) and CUB-200-2011 Wah et al. (2011). Specifically, we conduct experiments in a large-small collaboration scenario with heterogeneous multi-modality models, including two parties' SMs and one cloud-side MM-LM. Each party possesses both aligned and unaligned samples, where aligned samples share identical IDs but contain different textual descriptions, while unaligned samples are exclusively owned by individual parties. We evaluate three modality missing scenarios. In scenario 1, both parties and MM-LM suffer from text modality missing with probability Δ. In scenario 2, party 0 maintains complete image and text modalities while party 1 experiences text modality missing with probability Δ for both aligned and unaligned samples. In scenario 3, both parties and MM-LM suffer from image modality missing with probability Δ. Details of experimental datasets, settings and implementation details are provided in Appendix A.

Baselines: We compare our BoMM framework against five baseline methods: **1) LOCAL**, which trains local models using only local data without cloud-side collaboration; **2) MM-LOCAL**, which employs local multi-modal models to complete missing modalities before local training; **3) Vanilla** aggregates aligned data representations at the cloud-side MM-LM while parties train on their local data; **4) MM-AC**, which extends Vanilla by using the MM-LM to complete missing modalities in aligned data; and **5) MM-UC**, that further incorporates modality completion for both aligned and unaligned data at the cloud side.

5.2 Overall Performance

Tab. 1 presents the experimental results under both scenario 1 and scenario 2 with modality missing probability Δ set to 0.5. Our findings reveal that text modality completion provides limited performance gains for party's models, as evidenced by comparing MM-AC and MM-UC results where the inclusion of unaligned data completion shows minimal improvement. While this factor also constrains the improvement potential of our BoMM framework on local performance, our method still achieves consistent gains. Most notably, BoMM demonstrates substantial improvements on the cloud-side MM-LM performance. Specifically, under scenario 1 with zero aligned samples, BoMM achieves 4.75% improvement over the best baseline on MS COCO. Under scenario 2, the cloud-side improvement reaches 6.64% on MS COCO, highlighting the effectiveness of our bidirectional collaboration framework in enhancing domain-specific multi-modality large model capabilities.

Moreover, as shown in Tab. 2, scenario 3 demonstrates that image modality completion yields significantly greater performance improvements compared to text completion scenarios. This is evidenced by the substantial gains observed when comparing MM-UC against Vanilla, where Party 1's accuracy dramatically increases from 38.96% to 50.08%, representing an 11.12% improvement.

Table 1: Model performance comparison of BoMM under scenario 1 and scenario 2. P1 and P2 represent party 1 and party 2's performance, while C represents the cloud-side MM-LM performance.

Aligned Number	Method	Scenario 1						Scenario 2		
		MS COCO			CUB-200-2011			MS COCO		
		P1	P2	C	P1	P2	C	P1	P2	C
#0	LOCAL	38.67	38.28	-	56.34	29.98	-	50.48	34.96	-
	MM-LOCAL	36.13	36.52	-	57.62	31.84	-	50.68	36.52	-
	Vanilla	38.88	37.11	27.19	60.44	41.60	23.34	51.26	35.74	24.80
	MM-AC	39.06	37.40	27.64	59.45	46.38	23.27	50.20	35.84	26.26
	MM-UC	38.18	36.33	26.27	58.96	46.62	22.65	50.68	33.98	25.78
	BoMM	40.17	39.78	32.39	61.43	47.77	24.12	52.75	33.69	32.90
	Imp. (%)	1.99	2.38	4.75	0.99	1.15	0.78	0.49	-2.15	6.64
#200	LOCAL	39.47	37.41	-	60.85	37.25	-	50.98	37.41	-
	MM-LOCAL	39.14	36.10	-	60.36	35.58	-	51.97	36.10	-
	Vanilla	39.06	38.82	28.61	60.28	47.37	22.70	52.79	36.75	26.15
	MM-AC	38.15	39.80	28.45	60.03	43.25	23.11	51.94	36.84	25.65
	MM-UC	37.17	38.81	27.80	60.36	47.86	23.02	52.78	34.53	25.25
	BoMM	41.22	40.41	35.21	61.02	48.19	24.10	54.32	37.89	30.10
	Imp. (%)	1.75	0.61	6.60	0.17	0.82	0.99	1.53	1.05	3.95

Table 2: Model performance under scenario 3.

Method	MS COCO				
	P1		P2		C
	Acc	CLIPScore	Acc	CLIPScore	Acc
LOCAL	36.13	-	34.47	-	-
MM-LOCAL	30.18	-	36.62	-	-
Vanilla	38.96	-	38.18	-	27.83
MM-AC	40.79	64.12	39.88	65.67	28.04
MM-UC	50.08	62.16	39.72	67.56	44.90
BoMM	52.35	68.86	42.16	70.24	51.18
Imp. (%)	2.27	4.74	2.28	2.68	6.28

Building upon this foundation, our proposed BoMM framework achieves even more pronounced improvements on local performance, with Party 1 and Party 2 achieving additional 2.27% and 2.28% accuracy gains respectively over the strongest baseline MM-UC. The cloud-side MM-LM also benefits significantly from this collaborative approach, improving from 44.90% to 51.18%, demonstrating a 6.28% enhancement. Furthermore, we evaluate the quality of generated missing modalities using CLIPScore metrics Hessel et al. (2021), where BoMM consistently outperforms all baselines with scores of 68.86 and 70.24 for both parties, indicating superior generation quality that contributes to the overall model performance improvements.

5.3 Ablation Study

Table 3: Ablation study results.

Model	MS COCO			CUB-200-2011		
	P1	P2	C	P1	P2	C
BoMM	40.17	39.78	32.39	61.43	47.77	24.12
w/o preference loss	37.51	38.13	30.82	60.48	45.18	20.34
w/o special loss	38.47	35.35	30.66	60.05	40.23	7.23
w/o scheduler	38.37	35.15	30.56	59.96	41.50	11.13

We conduct ablation studies on scenario 2 with zero aligned samples to validate the contribution of key components in our BoMM framework. Specifically, removing the preference loss corresponds to setting $\mathcal{L}_{\text{prefer}}$ to 0, while removing the special loss involves setting the balancing hyperparameter γ to zero. For the scheduler ablation, we set the reward threshold μ to 0, which disables the sample filtering mechanism and includes all generated samples in training. The results in Tab. 3 demonstrate that each component contributes meaningfully to the overall performance of BoMM. Removing these components both lead to consistent performance degradation across both parties and the cloud-side MM-LM.

5.4 Convergence Experiments

Fig. 3(A) demonstrates the convergence behavior of our proposed BoMM framework compared with baseline methods across 5 training rounds. We evaluate convergence performance under the challenging zero-aligned samples scenario on both MS COCO and CUB-200-2011 datasets. Notably, BoMM exhibits faster convergence and reaches higher final accuracy compared to baseline methods, with both local parties and the MM-LM demonstrating steady performance improvements throughout the training process.

5.5 Hyperparameter Analysis

Fig. 3(B) presents the sensitivity analysis of two key hyperparameters in our BoMM framework. In Fig. 3(i), we evaluate the impact of modality missing rate Δ on performance across both datasets.

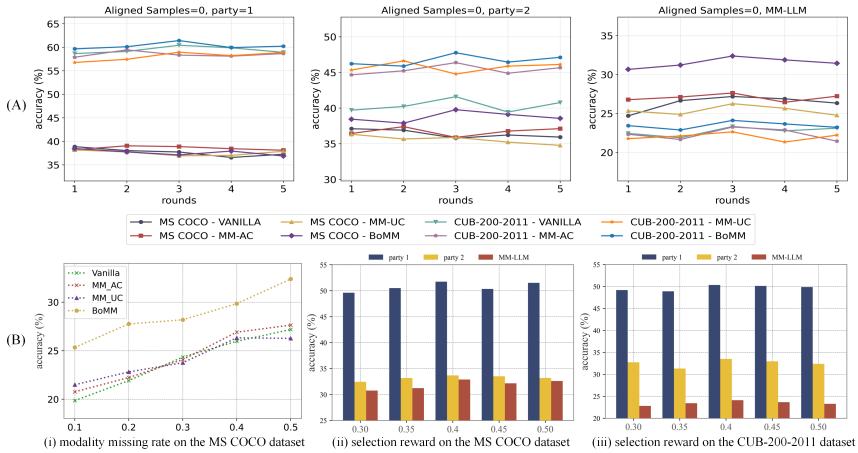

Figure 3: Convergence analysis and hyperparameter analysis of our proposed BoMM.

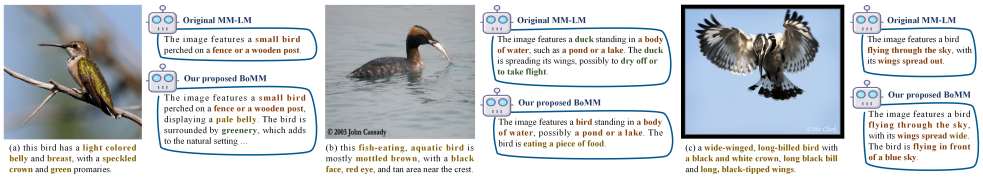

Figure 4: Case study results.

As expected, increasing modality missing rates lead to performance degradation, with our method maintaining stable performance even under severe missing conditions. Fig. 3(ii) and (iii) shows the effect of sample selection threshold μ on model performance. The results indicate that an optimal threshold around $\mu = 0.4$ achieves the best balance between sample quality and quantity, with too low thresholds including noisy samples and too high thresholds limiting data availability.

5.6 Case Study

The case study results in Fig. 4 compares text descriptions generated by the original MM-LM against our proposed BoMM method across three representative bird images. The color coding in the descriptions indicates different types of attributes: yellow bold text represents attributes that accurately match the dataset's standard descriptions, red bold text indicates attributes generated by the models that don't align with the dataset descriptions, and green bold text highlights obviously incorrect attributes. Specifically, BoMM consistently produces more detailed and contextually rich descriptions. For example, while the original method provides a basic description of "a small bird perched on a fence" our BoMM method enhances this with additional details such as "displaying a pale belly" and "surrounded by greenery". These improvements demonstrate that our preference-driven modality adaptive completion method successfully generates domain-specific, high-quality descriptions that better capture visual details.

6 Conclusion and Future Direction

This paper explores an under-explored paradigm: multi-modality large-small model collaboration. We propose a global prototype-guided alignment method to identify potentially aligned samples to mitigate modality alignment scarcity, and design a preference-driven modality adaptive completion method to address modality completeness gap challenges, enabling bidirectional large-small model knowledge transfer. Although we avoid direct leakage of raw data by exchanging intermediate representations, enhancing privacy remains an important future direction.

REFERENCES

Oded Berger-Tal, Jonathan Nathan, Ehud Meron, and David Saltz. The exploration-exploitation dilemma: a multidisciplinary framework. *PloS one*, 9(4):e95693, 2014.

Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. 2011.

Wanyi Chen, Zihua Zhao, Jiangchao Yao, Ya Zhang, Jiajun Bu, and Haishuai Wang. Multi-modal medical diagnosis via large-small model collaboration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30763–30773, 2025.

Sijie Cheng, Jingwen Wu, Yanghua Xiao, and Yang Liu. Fedgems: Federated learning of larger server models via selective knowledge fusion. *arXiv preprint arXiv:2110.11027*, 2021.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 958–979, 2024.

Tao Fan, Guoqiang Ma, Yan Kang, Hanlin Gu, Yuanfeng Song, Lixin Fan, Kai Chen, and Qiang Yang. Fedmkt: Federated mutual knowledge transfer for large and small language models. *arXiv preprint arXiv:2406.02224*, 2024.

Siwei Feng. Vertical federated learning-based feature selection with non-overlapping sample utilization. *Expert Systems with Applications*, 208:118097, 2022.

Siwei Feng, Boyang Li, Han Yu, Yang Liu, and Qiang Yang. Semi-supervised federated heterogeneous transfer learning. *Knowledge-Based Systems*, 252:109384, 2022.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Wei Guo, Fuzhen Zhuang, Xiao Zhang, Yiqi Tong, and Jin Dong. A comprehensive survey of federated transfer learning: challenges, methods and applications. *Frontiers of Computer Science*, 18(6):186356, 2024.

Wei Guo, Yiqi Tong, Yiyang Duan, Fuzhen Zhuang, Xiao Zhang, Zhaojun Hu, and Jin Dong. Feze: Alignment-flexible zero-shot vertical federated learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 745–754, 2025.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. Adaptive sample scheduling for direct preference optimization. *arXiv preprint arXiv:2506.17252*, 2025.

Kaiyi Ji, Jason D Lee, Yingbin Liang, and H Vincent Poor. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33:11490–11500, 2020.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024.

Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, Sergey Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation. *arXiv preprint arXiv:2410.18013*, 2024.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Soyeong Kwon, Taegyeong Lee, and Taehwan Kim. Zero-shot text-guided infinite image synthesis with llm guidance. *arXiv preprint arXiv:2407.12642*, 2024.

Iro Laina, Christian Rupprecht, and Nassir Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7414–7424, 2019.

Gyeonggeon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zhengliang Liu, Gengchen Mai, Tianming Liu, et al. Multimodality of ai for education: Towards artificial general intelligence. *IEEE Transactions on Learning Technologies*, 2025.

Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. *Advances in Neural Information Processing Systems*, 35:635–649, 2022.

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

Yang Liu, Bingjie Yan, Tianyuan Zou, Jianqing Zhang, Zixuan Gu, Jianbing Ding, Xidong Wang, Jingyi Li, Xiaozhou Ye, Ye Ouyang, et al. Towards harnessing the collaborative power of large and small models for domain tasks. *arXiv preprint arXiv:2504.17421*, 2025.

Yitao Liu, Tianxiang Sun, Xipeng Qiu, and Xuanjing Huang. Learning to teach with student feedback. *arXiv preprint arXiv:2109.04641*, 2021.

Zimo Liu, Kangjun Liu, Mingyue Guo, Shiliang Zhang, and Yaowei Wang. Cotuning: A large-small model collaborating distillation framework for better model generalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10487–10496, 2024.

Sahar Almahfouz Nasser, Nihar Gupte, and Amit Sethi. Reverse knowledge distillation: Training a large model using a small one for retinal image matching on limited data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7778–7787, 2024.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11557–11568, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

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

Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, and Weimin Zhang. How to bridge the gap between modalities: A comprehensive survey on multimodal large language model. *arXiv preprint arXiv:2311.07594*, 2023.

Yiqi Tong, Jiarui Zhang, Shaohang Wei, Wei Guo, Fuzhen Zhuang, Deqing Wang, Xi Yang, and Richeng Xuan. Mindscore: quantifying human preference for text-to-image generation through multi-view lens. *Science China Information Sciences*, 68(6):160105, 2025.

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

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*, 2024a.

Puyi Wang, Wei Sun, Zicheng Zhang, Jun Jia, Yanwei Jiang, Zhichao Zhang, Xiongkuo Min, and Guangtao Zhai. Large multi-modality model assisted ai-generated image quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7803–7812, 2024b.

Xiaosong Wang, Xiaofan Zhang, Guotai Wang, Junjun He, Zhongyu Li, Wentao Zhu, Yi Guo, Qi Dou, Xiaoxiao Li, Dequan Wang, et al. Openmedlab: An open-source platform for multi-modality foundation models in medicine. *arXiv preprint arXiv:2402.18028*, 2024c.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122. PMLR, 2015.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

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

Xun Wu, Shaohan Huang, Guolong Wang, Jing Xiong, and Furu Wei. Multimodal large language models make text-to-image generative models align better. *Advances in Neural Information Processing Systems*, 37:81287–81323, 2024b.

Zhengjie Yang, Sen Fu, Wei Bao, Dong Yuan, and Albert Y Zomaya. Fastslowmo: Federated learning with combined worker and aggregator momenta. *IEEE Transactions on Artificial Intelligence*, 4(5):1041–1050, 2022.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6137–6147, 2024.

Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023.

Haozhe Zhang, Junjie Pang, Yan Huang, Zhenzhen Xie, and Zelei Liu. Fedbridgeicl: Federated bridging of small and large models for in-context learning. In *International Conference on Wireless Artificial Intelligent Computing Systems and Applications*, pp. 1–10. Springer, 2025.

Huaao Zhang, Shigui Qiu, and Shilong Wu. Dual knowledge distillation for bidirectional neural machine translation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2021.

Weitong Zhang, Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural thompson sampling. *arXiv preprint arXiv:2010.00827*, 2020.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Bert learns to teach: Knowledge distillation with meta learning. *arXiv preprint arXiv:2106.04570*, 2021.

A Implement Details

We evaluate our proposed BoMM framework on two widely-used multi-modal datasets: MS COCO Lin et al. (2014) and CUB-200-2011 Wah et al. (2011). Specifically, the MS COCO dataset contains images across 80 object categories, where each image is paired with at least five textual descriptions. For our image-text classification task, we assign each image to the category with the highest number

of target objects present. The CUB-200-2011 dataset comprises 11,788 images covering 200 fine-grained bird species, with 5,994 images for training and 5,794 for testing. Following with Reed et al. (2016), we utilize the extended version that incorporates fine-grained visual descriptions, where each image contains at least 10 descriptions, with each description comprising at least 10 words without species names, background details, or action information.

We conduct experiments in a large-small collaboration scenario with heterogeneous multi-modality models, including two parties' SMs and one cloud-side MM-LM. The parties deploy small models including CLIP-base[1] and CLIP-large[2] locally, while the cloud hosts a large multi-modal model LLaVA-1.5-7B[3].Each party possesses both aligned and unaligned samples, where aligned samples share identical IDs but contain different textual descriptions, while unaligned samples are exclusively owned by individual parties. We evaluate three modality missing scenarios. In scenario 1, both parties and MM-LM suffer from text modality missing with probability Δ. In scenario 2, party 0 maintains complete image and text modalities while party 1 experiences text modality missing with probability Δ for both aligned and unaligned samples. In scenario 3, both parties and MM-LM suffer from image modality missing with probability Δ. In addition, to assess the impact of data alignment, we vary the number of aligned samples across 0 and 200, while fixing unaligned samples at 1,000 for each party and MM-LM.

We set the collaboration process to 3 global rounds, with a batch size of 32. In each round, both parties and the MM-LM perform 20 training epochs. All models are optimized using the Adam optimizer Kingma (2014), with learning rates set to 5e-4 for parties and 5e-5 for the MM-LM. Finally, we use accuracy to measure classification performance, CLIPScore Hessel et al. (2021) to evaluate the quality of generated modalities, communication cost measured in MB to assess data transfer efficiency, and computational time in minutes to evaluate system efficiency.

B Proof of Theorem 1

B.1 Assumptions

Let $\mathcal{Z} := \{\Psi, \Phi\}$ denote all model parameters. We adopt standard assumptions for bi-level optimization problem (Ghadimi & Wang, 2018; Ji et al., 2020) equation 10 on L_{global} and L_{local}.

Assumption 1. (Lipschitz Condition) The loss function $L_{\text{global}}(\mathcal{Z}; \xi)$ is L_1-Lipschitz for any given ξ.

Assumption 2. (Smoothness) The loss functions $L_{\text{global}}(\mathcal{Z}; \xi)$, $L_{\text{local}}(\mathcal{Z}; \xi)$ and $l_{\text{global}}(\mathcal{Z}; \xi)$ are L_2-smooth, L_2-smooth and L-smooth for any give ξ and ζ, respectively.

Assumption 3. (Lipschitz Condition for Second Derivatives) The second derivatives $\nabla_\Psi \nabla_\Theta L_{\text{local}}(\mathcal{Z}; \zeta)$ and $\nabla^2_\Theta L_{\text{local}}(\mathcal{Z}; \zeta)$ are L_3-Lipschitz and L_4-Lipschitz for any give ζ, respectively.

Assumption 4. (Bounded Domain) The parameter Θ is in a bounded domain with a diameter Δ, i.e., for any Θ_1 and Θ_2, we have:

$$\|\Theta_1 - \Theta_2\| \leq \Delta.$$

B.2 Abstraction of Our Algorithm

For clarity in convergence analysis, we write our algorithm as Algorithm 1.

B.3 Proof Details

To give the convergence analysis of our algorithm, we need to give some useful lemmas first.

Firstly, with L-smoothness of $l_{\text{local}}(\Psi, \Theta; \zeta)$, we could give the strongly convexity property of $L_{\text{local}}(\Psi, \Theta; \zeta)$.

[1] https://huggingface.co/openai/clip-vit-base-patch32

[2] https://huggingface.co/openai/clip-vit-large-patch14

[3] https://huggingface.co/llava-hf/llava-1.5-7b-hf

Algorithm 1 Algorithmic Abstraction of BoMM

Input: communication round T; learning rate: η and η'; epoch: τ, τ'; initial parameters Θ_{-1}^{τ} and $\Psi_{-1}^{\tau'}$.
for $t = 0, 1, \ldots, T-1$ **do**
  Let $\Theta_t^0 = \Theta_{t-1}^{\tau}$ and $\Psi_t^0 = \Psi_{t-1}^{\tau'}$;
  for $j' = 1, 2, \ldots, \tau'$ **do**
    Select S'_{j-1}, a subset of each party's dataset D_k,
    update $\Psi_t^{j'} = \Psi_t^{j'-1} - \eta' \nabla_{\Psi} L_{\text{global}}(\Psi_t^{j'-1}, \Theta_{t-1}^{\tau}; S'_{j-1})$;
  end for
  for $j = 1, 2, \ldots, \tau$ **do**
    Update $\Theta_t^j = \Theta_t^{j-1} - \eta \nabla_{\Theta} L_{\text{local}}(\Psi_t^{\tau'}, \Theta_t^{j-1})$;
  end for
end for

Lemma 1. Under Assumption 2, suppose $\alpha := -L + \varphi > 0$, $L_{\text{local}}(\Psi, \Theta; \zeta)$ is α-strongly convex w.r.t Θ.

Based on Assumptions 1 and 2, we could give Lemma 2 directly.

Lemma 2. Under Assumption 1 and 2, the derivatives $\nabla L_{\text{global}}(\mathcal{Z}; \xi)$, $\nabla_{\Psi} \nabla_{\Theta} L_{\text{local}}(\mathcal{Z}; \zeta)$ and $\nabla_{\Theta}^2 L_{\text{local}}(\mathcal{Z}; \zeta)$ have bounded variances, i,e., for any $\mathcal{Z}$, we have

$$\mathbb{E}_{\xi} \|\nabla L_{\text{global}}(\mathcal{Z}; \xi) - \nabla \mathcal{L}_{\text{global}}(\mathcal{Z})\|^2 \leq L_1^2, \tag{12}$$

$$\mathbb{E}_{\zeta} \|\nabla_{\Psi} \nabla_{\Theta} L_{\text{local}}(\mathcal{Z}; \zeta) - \nabla_{\Psi} \nabla_{\Theta} \mathcal{L}_{\text{local}}(\mathcal{Z})\|^2 \leq L_2^2, \tag{13}$$

$$\mathbb{E}_{\zeta} \|\nabla_{\Theta}^2 L_{\text{local}}(\mathcal{Z}; \zeta) - \nabla_{\Theta}^2 \mathcal{L}_{\text{local}}(\mathcal{Z})\|^2 \leq L_2^2. \tag{14}$$

We omit them since the proof of Lemma 1 and 2 is basic.

To show the smooth property of $F(\Psi)$, we introduce the following lemma which is proposed in Ghadimi & Wang (2018) firstly.

Lemma 3. (Lemma 2.2 in Ghadimi & Wang (2018)) Under Assumptions 1, 2 and 3, $F(\Psi)$ is L_0-smooth where L_0 is given by

$$L_0 := L_2 + \frac{2L_2^2 + L_1^2 L_3}{\alpha} + \frac{L_1 L_2 L_3 + L_1 L_2 L_4 + L_2^3}{\alpha^2} + \frac{L_1 L_2^2 L_4}{\alpha^3}.$$

Tracking error $\mathbb{E}\|\Theta_t^{j-1} - \Theta^*(\Psi_t^0)\|$ is an important component in our convergence analysis. To give an upper bound on the tracking error, we utilized Lemma 9 in Ji et al. (2021). For simplicity, we assume equal batch sizes N_b for both cloud and local data.

Lemma 4. (Lemma 9 in Ji et al. (2021)) Under Assumptions 1 and 2, with stepsize η to be $\frac{2}{L_2+\alpha}$, we have

$$\mathbb{E}\|\Theta_t^{j-1} - \Theta^*(\Psi_t^0)\|^2 \leq \left(\frac{L_2 - \alpha}{L_2 + \alpha}\right)^{2(j-1)} \mathbb{E}\|\Theta_t^0 - \Theta^*(\Psi_t^0)\|^2. \tag{15}$$

With the help of the above lemmas, we could give the estimation property of the $\frac{\partial L_{\text{global}}(\Psi_t^0, \Theta_t^{\tau})}{\partial \Psi_t^0}$ approximating $\nabla F(\Psi_t^0)$. The result is presented in the following Proposition 1.

Proposition 1. Under Assumptions 1-4, choose stepsize η to be $\frac{2}{L_2+\alpha}$ and suppose $\alpha < L_2$, we have

$$\begin{aligned}\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^0,\Theta_t^\tau;\mathcal{S}_0')}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\| \le (L_2+\frac{L_2^2}{\alpha})[\Big(\frac{L_2-\alpha}{L_2+\alpha}\Big)^\tau \\ \sqrt{\Delta}]+L_1[\frac{L_2(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{\alpha} \\ +\frac{2}{L_2+\alpha}(\frac{L_2L_4}{\alpha}+L_3)\sqrt{\Delta} \\ \frac{(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{1-\frac{2}{L_2+\alpha}\alpha-\frac{L_2-\alpha}{L_2+\alpha}}]+\frac{L_1}{\sqrt{N_b}}.\end{aligned} \tag{16}$$

Proof. Using the triangle inequality, we have

$$\begin{aligned}&\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^0,\Theta_t^\tau;\mathcal{S}_0')}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\| \\ &=\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^0,\Theta_t^\tau;\mathcal{S}_0')}{\partial\Psi_t^0}-\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0} \\ &+\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\| \\ &\le\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^0,\Theta_t^\tau;\mathcal{S}_0')}{\partial\Psi_t^0}-\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}\| \\ &+\mathbb{E}\|\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\| \\ &\overset{(i)}{\le}\frac{L_1}{\sqrt{N_b}}+\mathbb{E}\|\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\|.\end{aligned} \tag{17}$$

where (i) follows from equation 12.

Then, we need to give an upper bound for $\mathbb{E}\|\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\|$. Using

$$\nabla F(\Psi_t^0)=\nabla\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta^*(\Psi_t^0))+\frac{\partial\Theta^*(\Psi_t^0)}{\partial\Psi_t^0}\nabla_\Theta\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta^*(\Psi_t^0))$$

and

$$\frac{\partial L_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}=\nabla_\Psi\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)+\frac{\partial\Theta_t^\tau}{\partial\Psi_t^0}\nabla_\Theta\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau),$$

we have

$$\begin{aligned}&\mathbb{E}\|\frac{\partial\mathcal{L}_{\text{global}}(\Psi_t^0,\Theta_t^\tau)}{\partial\Psi_t^0}-\nabla F(\Psi_t^0)\| \\ &\le L_2\mathbb{E}\|\Theta_t^\tau-\Theta^*(\Psi_t^0)\|+L_1\mathbb{E}\|\frac{\partial\Theta_t^\tau}{\partial\Psi_t^0}-\frac{\partial\Theta^*(\Psi_t^0)}{\partial\Psi_t^0}\| \\ &+L_2\mathbb{E}(\|\frac{\partial\Theta^*(\Psi_t^0)}{\partial\Psi_t^0}\|\|\Theta_t^\tau-\Theta^*(\Psi_t^0)\|).\end{aligned} \tag{18}$$

Now we want to bound $\mathbb{E}\|\frac{\partial\Theta_t^\tau}{\partial\Psi_t^0}-\frac{\partial\Theta^*(\Psi_t^0)}{\partial\Psi_t^0}\|$ first.

Recall the update method

$$\Theta_t^j=\Theta_t^{j-1}-\eta\nabla_\Theta\mathcal{L}_{\text{local}}(\Psi_t^0,\Theta_t^{j-1})$$

and use the chain rule on it, we have

$$\begin{aligned}\frac{\partial\Theta_t^j}{\partial\Psi_t^0}=\frac{\partial\Theta_t^{j-1}}{\partial\Psi_t^0}-\eta(\nabla_\Psi\nabla_\Theta\mathcal{L}_{\text{local}}(\Psi_t^0,\Theta_t^{j-1}) \\ +\frac{\partial\Theta_t^{j-1}}{\partial\Psi_t^0}\nabla_\Theta^2\mathcal{L}_{\text{local}}(\Psi_t^0,\Theta_t^{j-1})).\end{aligned} \tag{19}$$

For $\Theta^*(\Psi_t^0)$ is the optimal solution of $\mathcal{L}_{\text{local}}(\Psi_t^0, \Theta)$, we have $\nabla_\Theta \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta^*(\Psi_t^0)) = 0$. Then, using the chain rule, we have

$$\nabla_\Psi \nabla_\Theta \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta^*(\Psi_t^0)) + \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta^*(\Psi_t^0)) = 0. \tag{20}$$

Combining equation 19 and equation 20, the following equation holds

$$\begin{aligned}
\frac{\partial \Theta_t^j}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} &= \frac{\partial \Theta_t^{j-1}}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \\
&- \eta \left(\frac{\partial \Theta_t^{j-1}}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \right) \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^{j-1}) \\
&- \eta \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} (\nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^{j-1}) - \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta^*(\Psi_t^0))).
\end{aligned} \tag{21}$$

Based on equation 20, we have

$$\| \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \| \leq \frac{L_2}{\alpha}. \tag{22}$$

With the help of Assumption 3, Lemma 2, equation 21 and equation 22, the following bound holds

$$\begin{aligned}
&\mathbb{E} \| \frac{\partial \Theta_t^j}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \| \\
&\leq \mathbb{E}(\| I - \eta \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^{j-1}) \| \| \frac{\partial \Theta_t^{j-1}}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \|) \\
&+ \eta \frac{L_2}{\alpha} (\mathbb{E} \| \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^{j-1}) - \nabla_\Theta^2 \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta^*(\Psi_t^0)) \| \\
&+ \eta \mathbb{E} \| \nabla_\Psi \nabla_\Theta \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^{j-1}) - \nabla_\Psi \nabla_\Theta \mathcal{L}_{\text{local}}(\Psi_t^0, \Theta_t^*(\Psi_t^0)) \| \\
&\leq (1 - \eta\alpha) \mathbb{E} \| \frac{\partial \Theta_t^{j-1}}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \| + \eta (\frac{L_2 L_4}{\alpha} + L_3) \mathbb{E} \| \Theta_t^{j-1} - \Theta^*(\Psi_t^0) \|.
\end{aligned} \tag{23}$$

For the choice of $\eta = \frac{2}{L_2 + \alpha}$, Lemma 4 holds. With the result in equation 15 and Assumption 4, we have

$$\begin{aligned}
\mathbb{E} \| \Theta_t^{j-1} - \Theta^*(\Psi_t^0) \| &\leq \left(\frac{L_2 - \alpha}{L_2 + \alpha} \right)^{j-1} \sqrt{\mathbb{E} \| \Theta_t^0 - \Theta^*(\Psi_t^0) \|} \\
&\leq \left(\frac{L_2 - \alpha}{L_2 + \alpha} \right)^{j-1} \sqrt{\Delta}.
\end{aligned} \tag{24}$$

Telescoping equation 23 over j from 0 to τ and combining the result with equation 24 yields

$$\begin{aligned}
\mathbb{E} \| \frac{\partial \Theta_t^\tau}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \| &\leq (1 - \eta\alpha)^\tau \mathbb{E} \| \frac{\partial \Theta_t^0}{\partial \Psi_t^0} - \frac{\partial \Theta^*(\Psi_t^0)}{\partial \Psi_t^0} \| \\
&+ \eta \left(\frac{L_2 L_4}{\alpha} + L_3 \right) \sum_{j=0}^{\tau-1} (1 - \eta\alpha)^{\tau-1-j} \\
&\left[\left(\frac{L_2 - \alpha}{L_2 + \alpha} \right)^j \sqrt{\Delta} \right] \\
&\leq \frac{L_2 (1 - \eta\alpha)^\tau}{\alpha} \\
&+ \eta (\frac{L_2 L_4}{\alpha} + L_3) \sqrt{\Delta} \frac{(1 - \eta\alpha)^\tau}{1 - \eta\alpha - \frac{L_2 - \alpha}{L_2 + \alpha}}.
\end{aligned} \tag{25}$$

Take expedition of both sides of equation 18, plugging equation 22, equation 24 and equation 25 into it yields

$$\begin{aligned}\mathbb{E}\|\frac{\partial \mathcal{L}_{\text{global}}(\Psi_t^0, \Theta_t^\tau)}{\partial \Psi_t^0} - \nabla F(\Psi_t^0)\| \leq (L_2 + \frac{L_2^2}{\alpha})[\left(\frac{L_2-\alpha}{L_2+\alpha}\right)^\tau \sqrt{\Delta}] \\ + L_1[\frac{L_2(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{\alpha} \\ + \frac{2}{L_2+\alpha}(\frac{L_2L_4}{\alpha} + L_3)\sqrt{\Delta} \\ \frac{(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{1-\frac{2}{L_2+\alpha}\alpha - \frac{L_2-\alpha}{L_2+\alpha}}].\end{aligned} \tag{26}$$

Finally, plugging equation 26 into equation 17 yields equation 16. Thus, we complete the proof of Propostion 1. □

To deal with multiple updates of the active party with a fixed t, we need to use a technology called virtual updates (Yang et al., 2022). Specifically, we introduce a virtual parameter $\Theta_{t,j'}^\tau$ which could be obtained by local updates when $\Psi_t^{j'}$ is given. With the help of Proposition 1, we could obtain the bound for $\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|$. Now, we want to give a bound for $\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|$. The result is as follows

Proposition 2. Following the conditions in Proposition 1, we have

$$\begin{aligned}\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\| \leq (L_2 + \frac{L_2^2}{\alpha})[\left(\frac{L_2-\alpha}{L_2+\alpha}\right)^\tau \\ \sqrt{\Delta}] + L_1[\frac{L_2(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{\alpha} \\ + \frac{2}{L_2+\alpha}(\frac{L_2L_4}{\alpha} + L_3)\sqrt{\Delta} \\ \frac{(1-\frac{2}{L_2+\alpha}\alpha)^\tau}{1-\frac{2}{L_2+\alpha}\alpha - \frac{L_2-\alpha}{L_2+\alpha}}] + \frac{L_1}{\sqrt{N_b}} + L_2\Delta.\end{aligned} \tag{27}$$

Proof. Using the triangle inequality, we have

$$\begin{aligned}&\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\| \\ &\leq \|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}}\| \\ &+ \|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|\end{aligned} \tag{28}$$

Take expectation of both sides of equation 28, with the help of Assumption 2 and 5, the following result holds

$$\begin{aligned}
&\mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\| \\
&\leq \mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}}\| \\
&+ \mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\| \\
&\leq L_2\|\Theta_t^\tau - \Theta_{t,j'}^\tau\| + \mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\| \\
&\leq L_2\Delta + \mathbb{E}\|\frac{\partial L_{\text{global}}(\Psi_t^{j'}, \Theta_{t,j'}^\tau; \mathcal{S}'_{j'})}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|.
\end{aligned} \tag{29}$$

Plugging in the result in Proposition 1 into equation 29 yields equation 27. Thus, we complete the proof of Propostion 2. □

Proof for Theorem 1. Based on the L_0-smoothness of $F(\Psi)$ established in Lemma 3, we have

$$\begin{aligned}
F(\Psi_t^{j'+1}) &\leq F(\Psi_t^{j'}) + \langle \nabla F(\Psi_t^{j'}), \Psi_t^{j'+1} - \Psi_t^{j'} \rangle \\
&+ \frac{L_0}{2}\|\Psi_t^{j'+1} - \Psi_t^{j'}\|^2 \\
&\leq F(\Psi_t^{j'}) - \eta' \langle \nabla F(\Psi_t^{j'}), \frac{\partial \mathcal{L}_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau)}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'}) \rangle \\
&- \eta' \|\nabla F(\Psi_t^{j'})\|^2 + \eta'^2 L_0 \|\nabla F(\Psi_t^{j'})\|^2 \\
&+ \eta'^2 L_0 \|\frac{\partial \mathcal{L}_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau)}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|^2 \\
&\leq F((\Psi_t^{j'}) - (\frac{\eta'}{2} - \eta'^2 L_0)\|\nabla F((\Psi_t^{j'})\|^2 \\
&+ (\frac{\eta'}{2} + \eta'^2 L_0)\|\frac{\partial \mathcal{L}_{\text{global}}(\Psi_t^{j'}, \Theta_t^\tau)}{\partial \Psi_t^{j'}} - \nabla F(\Psi_t^{j'})\|^2.
\end{aligned} \tag{30}$$

Telescoping equation 30 over j' from 0 to $\tau' - 1$ and taking expection of both sides yields

$$\begin{aligned}
\mathbb{E}F(\Psi_t^{\tau'}) \leq \mathbb{E}F(\Psi_t^0) &- (\frac{\eta'}{2} - \eta'^2 L_0)\mathbb{E}\sum_{j'=0}^{\tau'-1} \|\nabla F(\Psi_t^{j'})\|^2 \\
&+ (\eta' + 2\eta'^2 L_0)\tau'\{(L_2 + \frac{L_2^2}{\alpha})[\left(\frac{L_2 - \alpha}{L_2 + \alpha}\right)^\tau \\
&\sqrt{\Delta}] + L_1[\frac{L_2(1 - \frac{2}{L_2+\alpha}\alpha)^\tau}{\alpha} \\
&+ \frac{2}{L_2 + \alpha}(\frac{L_2 L_4}{\alpha} + L_3)\sqrt{\Delta} \\
&\frac{(1 - \frac{2}{L_2+\alpha}\alpha)^\tau}{1 - \frac{2}{L_2+\alpha}\alpha - \frac{L_2-\alpha}{L_2+\alpha}}] + \frac{L_1}{\sqrt{N_b}}\}^2 \\
&+ (\eta' + 2\eta'^2 L_0)L_2^2\Delta^2(\tau' - 1).
\end{aligned} \tag{31}$$

Telescoping equation 31 over t from 0 to $T-1$ yields equation 11, we complete the proof of Theorem 1. □