Reviewer 1:

Question 1: The motivation for transitioning from FedReFT to FedReFT+ is not clearly articulated, making it difficult to understand the limitations of FedReFT that FedReFT+ aims to overcome. As described, FedReFT resembles a standard FedAvg approach where clients apply ReFT locally and the server aggregates the trainable parameters via weighted averaging, yet the paper does not clarify why this baseline is insufficient.

Answer: We appreciate the reviewer's thoughtful feedback. There is no prior method called "FedReFT" upon which our work builds. Our proposed FedReFT+ is not an incremental improvement over an existing "FedReFT" method, but rather a new framework that introduces Representation Fine-Tuning (ReFT) into the federated learning (FL) setting for the first time, to the best of our knowledge. The name "FedReFT+" reflects our broader contributions beyond simply adapting ReFT to FL. Specifically: We identify that naively aggregating representation-level updates across heterogeneous clients can lead to semantic misalignment, especially in task-diverse FL settings. To address this, we propose a novel All-But-Me (ABM) aggregation strategy that improves semantic stability by letting clients partially incorporate global representation shifts without diluting their local semantics.

Question 2: The adoption of the geometric median in the All-but-Me aggregation strategy lacks strong justification. While an ablation study is provided in Appendix F, the performance differences between aggregation methods are marginal, and the geometric median introduces higher computational complexity compared to simpler alternatives like Mean-ABM.

Answer: While the performance gap may appear small at first glance, we would like to emphasize that even modest improvements are meaningful in the FL setting, particularly when evaluating under highly heterogeneous task distributions. As shown in Figure 3 (Appendix F) and Table 1: The time complexity for the arithmetic mean is O(d), whereas the time complexity for the geometric median (using Weiszfeld's algorithm) is O(T.d), where T is the number of iterations and d is the number of trainable parameters. For memory complexity, both the arithmetic mean and geometric median have the same complexity of O(d). Despite being computationally more expensive, the geometric median is more robust for heterogeneous aggregation and often yields higher accuracy in federated learning (FL) settings. The following table shows the performance of FedAvg, Mean_ABM, and GeoMedian_ABM:

Task	Method	Accuracy (%)	Accu Δ (GeoMed vs. others)	Params (M)
Commonsense,LLaMa-2 7B	FedAvg	70.26	+0.73%	4.70
	Mean_ABM	70.58	+0.41%	4.70
	GeoMedian_ABM	70.99	–	4.70
Arithmetic,LLaMa-2 7B	FedAvg	15.35	+1.86%	4.70
	Mean_ABM	16.20	+1.01%	4.70
	GeoMedian_ABM	17.21	–	4.70
GLUE,RoBERTa	FedAvg	51.30	+1.06%	0.053

Task	Method	Accuracy (%)	Accu Δ (GeoMed vs. others)	Params (M)
	Mean_ABM	52.17	+0.19%	0.053
	GeoMedian_ABM	52.36	–	0.053

These improvements are consistent across three diverse task types. We chose Geometric Median-ABM not for peak accuracy alone but for its robustness to outlier updates and task drift, which are common in federated personalization scenarios. Especially in the arithmetic reasoning task, which is highly sensitive to client diversity, the gains were more pronounced. While we acknowledge the additional cost of computing the geometric median, this is done only once per round at the server on low-dimensional sparse intervention parameters (not full models), making the overhead negligible. Importantly, client-side efficiency is unaffected, preserving our design goal of being lightweight for edge devices. We will revise Appendix F to include variance/error bars and improve clarity in the camera-ready version.

Question 3: The proposed FedReFT+ does not consistently outperform existing methods across all benchmarks. In both Table 3 and Table 5, several baselines achieve better performance, raising concerns about the practical utility and competitiveness of FedReFT+ in real-world applications despite the authors' efforts to improve performance.

Answer: The centralized standalone ReFT baseline (Wu et al., 2024b) does not consistently outperform other PEFT methods in accuracy, whereas it is 15–65× more parameter-efficient than LoRA while still achieving competitive accuracy. Therefore, FedReFT+ delivers the best balance of Parameter efficiency and performance. We acknowledge that FedReFT+ may not achieve the highest score on every benchmark. However, when considering accuracy and parameter efficiency, FedReFT+ nearly outperforms all state-of-the-art PEFT methods in Federated Learning settings. The following tables show how much FedReFT+ is efficient compared to the SOTA approaches.

From Table 3, Federated fine-tuning performance of LlaMa-3.2 3B across five commonsense reasoning tasks with Mixed Task (MT) experimental setup, where clients train on heterogeneous task mixtures to promote generalizable representations.

Method	Rank (R)	Param (M)	Avg Accu (%)	FedReFT+ (R 32) Param Effi.	Accu Δ (FedReFT+(R 32) vs. others)	FedReFT+ (R 8) Param Effi.	Accu Δ (FedReFT+(R 8) vs. others)
FLoRA	32	243.15	78.83	22.09×	−2.61%	88.42×	−3.17%
FedIT	32	48.63	75.74	4.42×	+0.48%	17.68×	−0.08%
FFA-LoRA	32	24.31	71.11	2.21×	+5.11%	8.84×	+4.55%
Fed-SB	120	2.83	75.66	0.26×	+0.56%	1.03×	0.00%
FedReFT+	32	11.01	76.22	—	—	4.00×	+0.56%
FedReFT+	8	2.75	75.66	0.25×	−0.56%	—	—

From Table 5, Performance comparison across GLUE Tasks on RoBERTa model for C = 3, FedReFT+ uses rank 1, whereas all baseline uses rank 8. For the Natural Language Understanding task, FedReFT+ outperforms all the baselines and is (27.17× to 34.53×) times more parameter efficient.

Method	Trainable Param (M)	Avg Accu (%)	FedReFT+ Param Effi.	FedReFT+ Accu (±)
FFA-LoRA	1.44	89.39	27.17×	+1.54%
FedDPA-LoRA	2.62	89.47	49.43×	+1.46%
FedSA-LoRA	1.83	90.43	34.53×	+0.50%
FedReFT+	0.053	90.93	-	-

From Table 6, Performance comparison on arithmetic reasoning tasks for GSM8K on LLaMa-3 8B model with LoRA rank 8, where clients enable consistent evaluation of representation generalization. FedReFT+ achieves (+3.05%, +3.36%) higher accuracy with (3.63×, 7.25×) times parameter efficient.

Method	Rank	Trainable Param (M)	Accu (%)	FedReFT+ Param Effi.	FedReFT+ Accu (±)
FedSA-LoRA	8	30.40	46.63	7.25×	+3.05%
FFA-LoRA	8	15.20	46.32	3.63×	+3.36%
FedReFT+	8	4.19	49.68	-	-

The tie-ϕ variant further demonstrates how our method scales down to even more compact configurations with acceptable performance loss.

Reviewer 2

Question 1: The authors claim that FedAvg can cause semantic interference or collapse for LoReFT in FL settings. However, the claim of collapsing is not supported by analysis or experiments.

Answer: We appreciate the reviewer highlighting this point. ReFT methods operate on a frozen base model and learn task-specific Interventions on hidden representations. Our claim regarding semantic interference or collapse under FedAvg is grounded in the intuitive mismatch between LoReFT-style updates (ReFT, LoRA, LoReFT) and naive intervention parameter averaging, especially under heterogeneous client tasks. Since LoReFT modifies internal representation layers, averaging such updates across clients with divergent tasks can result in semantic drift, where the aggregated representation no longer aligns with any client’s local task semantics.

As shown in Figure 3 (Appendix F) and Table 1: The time complexity for the arithmetic mean is O(d), whereas the time complexity for the geometric median (using Weiszfeld’s algorithm) is O(T.d), where T is the number of iterations and d is the number of trainable parameters. For memory complexity, both the arithmetic mean and geometric median have the same complexity of O(d). Despite being computationally more expensive, the geometric median is more robust for heterogeneous aggregation and often yields higher accuracy in federated learning (FL) settings. The following table shows the performance of FedAvg, Mean_ABM, and GeoMedian_ABM:

Task	Method	Accuracy (%)	Accu Δ (GeoMed_ABM vs. others)	Params (M)
Commonsense,LLaMa-2 7B	FedAvg	70.26	+0.73%	4.70
	Mean_ABM	70.58	+0.41%	4.70
	GeoMedian_ABM	70.99	–	4.70
Arithmetic,LLaMa-2 7B	FedAvg	15.35	+1.86%	4.70
	Mean_ABM	16.20	+1.01%	4.70
	GeoMedian_ABM	17.21	–	4.70
GLUE,RoBERTa	FedAvg	51.30	+1.06%	0.053
	Mean_ABM	52.17	+0.19%	0.053
	GeoMedian_ABM	52.36	–	0.053

Question 2: The authors claim that FedAvg can cause semantic interference or collapse for LoReFT in FL settings. However, the claim of collapsing is not supported by analysis or experiments. All the discussions about existing literature and challenges talk about the potential incompatibility between FedAvg and LoReFT. In the experiments, the comparisons are made against FLoRA, FedIT, FFA-LoRA, FedSB, etc. None of these methods is analyzed or discussed to show the methodologies used by these methods and how these methods would solve the aforementioned problems. Given the names of these methods, it could be that these methods are not related to LoReFT. In this case, the authors should discuss/show what happens if the FedAvg and LoReFT are directly combined and implemented. In its current form, the research questions brought up in the introduction are left unanswered.

Answer: We sincerely appreciate the reviewer’s thoughtful comment and agree that a more direct analysis of FedAvg+LoReFT compatibility would strengthen the clarity of our contributions. We clarify the following in response:

FedAvg on LoReFT is not well-established in the existing literature. For any Federated Learning (FL) setting, FedAvg is the vanilla aggregation method. So we first tried FedAvg with LoReFT in FL, but (as mentioned in the previous question's answer), since it is a representation fine-tuning method, we aimed to preserve the personalized representation fine-tuning locally at each client, rather than simply averaging on the server side. Our motivation arises from the hypothesis—grounded in empirical evidence and architectural intuition—that naive aggregation of representation-level interventions (via FedAvg) can lead to semantic drift or collapse, particularly in heterogeneous tasks.

Regarding baseline selection: While FLoRA, FedIT, FFA-LoRA, and FedSB are not explicitly LoReFT-based, they are the closest state-of-the-art parameter-efficient FL methods, often relying on LoRA-like decompositions. We selected them intentionally to benchmark our method against the strongest available alternatives in low-rank or PEFT-based FL.

On FedAvg + LoReFT baseline: An explicit implementation of a naïve FedAvg+LoReFT combination is shown in Appendix F.1, Figure 3. This represents the FedAvg approach applied to LoReFT in FL settings, which we refer to as LoReFT+FedAvg under the FedReFT+ framework. Additionally, Table 1 provides a tabular view of the same setting without our ABM aggregation.

We will also enhance the related work section to clarify which baselines do or do not use representation-level interventions, and how their aggregation strategies differ. Thank you again for pointing this out—we will revise accordingly in the final version.

Question 3: The W2 could be partly answered by the contents from Appendix F.1. However, Fig.3 in Appendix F.1 shows that the proposed method (ABM) provides marginal improvements to FedAvg. In addition, the results shown by Fig.3 in Appendix F.1 are not consistent with the results in the main context. No parameters about the used models are given for Appendix F.1.

Answer: We did the experiments on a reduced dataset for this section, so the results are not the same as the main context. We present the same data in Table 1, which shows the result of the GLUE task on ROBERTa, Commonsense, and Arithmetic reasoning task on LLaMA-2 7B for three clients.

Task	**Method**	**Accuracy (%)**	**Accu Δ (GeoMed_ABM vs. others)**	**Params (M)**
Commonsense,LLaMa-2 7B	FedAvg	70.26	+0.73%	4.70
	Mean_ABM	70.58	+0.41%	4.70
	GeoMedian_ABM	70.99	–	4.70
Arithmetic,LLaMa-2 7B	FedAvg	15.35	+1.86%	4.70
	Mean_ABM	16.20	+1.01%	4.70
	GeoMedian_ABM	17.21	–	4.70
GLUE,RoBERTa	FedAvg	51.30	+1.06%	0.053
	Mean_ABM	52.17	+0.19%	0.053
	GeoMedian_ABM	52.36	–	0.053

Question 4: The heterogeneous distribution among clients may not always occur, although it could appear in some real-world FL applications. The Distinct Task (DT) scenario is closer to the heterogeneous distribution assumption. However, no DT results are given for 3.1 Commonsense Reasoning; only Mixed Task (MT) results are shown in Table 3. There is no description about DT/MT for 3.3 Natural Language Understanding. Table 4 gives DT results for 3.2 Arithmetic Reasoning; however, (1) models perform better in DT rather than MT; (2) no comparisons are made to other baselines.

Answer: We conducted commonsense reasoning experiments under the Distinct Task (DT) and Mixed Task (MT) setups, as described in the paper. Due to space constraints, we moved the MT experiments to Appendix G.1 (Table 14). While the caption of Table 14 does not explicitly mention the commonsense reasoning task, we clarify that the experiments presented there are indeed conducted on the commonsense reasoning dataset.

- There is no description about DT/MT for 3.3 Natural Language Understanding Answer: For the Natural Language Understanding (NLU) experiments, we used the GLUE benchmark. However, these experiments do not follow the DT/MT task distribution paradigm. Instead, we partitioned each GLUE task among clients, allowing local training on task-specific subsets. Each client was evaluated using its local test set from the same task. We applied this experimental design uniformly across all GLUE datasets. Since this setup does not involve cross-task distribution (DT/MT), we did not include it under those sections. The primary objective was to examine whether lightweight intervention tuning can effectively align representations across clients within a single NLU task.

- Table 4 gives DT results for 3.2 Arithmetic Reasoning; however, (1) models perform better in DT rather than MT; Answer: In the Mixed Task (MT) setting, each client trains on a subset of a combined reasoning dataset, encouraging generalization across tasks. Conversely, in the Distinct Task (DT) setting, each client trains on a unique reasoning task, enabling more personalized fine-tuning while benefiting from global updates. As expected, the DT setup typically results in higher accuracy, as clients specialize in a single task.
- (2) No comparisons are made to other baselines in Table 4. Answer: Regarding comparisons in Table 5, to the best of our knowledge, no existing works follow the DT/MT experimental design over math10k, which limited our ability to include direct comparisons in that table. However, we referenced several relevant works on arithmetic reasoning tasks on the GSM8K dataset and included comparative results in Table 6. FedReFT+ achieved over 3+% accuracy gain compared to the SOTA.

Question 5: The related work section is placed in the appendix to bypass the page limit. Some content in the Appendices is important to the paper and should also be included in the main context, such as some portion of the related work, the analysis of ABM, and the ablation study.

Answer: Thank you for pointing this out. Due to page limitations, we moved the Related Work, ABM analysis, and ablation study to the appendix to prioritize core content in the main paper. If accepted, we will include concise versions of these important sections in the camera-ready version to improve clarity and completeness.

Question 6: In the author's checklist, the authors have checked B4 PII and Offensive information, but no required elaboration/explanation is given in the annotated Section 3, nor in the whole paper.

Answer: We sincerely apologize for the oversight; our work uses only publicly available, licensed datasets that do not contain any B4 PII or offensive content. We will correct the mistaken selection of item B4 in the camera-ready version.

Reviewer 3

Question 1: In Table 3, FedReFT+ with rank=8 shows no improvement over Fed-SB in terms of both performance and the number of trainable parameters.

Answer: The following table shows the performance efficiency and accuracy status of FedReFT+ over other baselines. FedReFT+ with rank 8 achieves a similar accuracy of 75.66%, while being 1.03× more parameter-efficient compared to Fed-SB. In contrast, Fed-SB utilizes a much higher LoRA rank of 120, which significantly increases the number of trainable parameters and contributes to its comparable accuracy. From Table 3, Federated fine-tuning performance of LlaMa-3.2 3B across five commonsense reasoning tasks with Mixed Task (MT) experimental setup, where clients train on heterogeneous task mixtures to promote generalizable representations.

Method	Rank (R)	Param (M)	Avg Accu (%)	FedReFT+ (R 32) Param Effi.	Accu Δ (FedReFT+(R 32) vs. others)	FedReFT+ (R 8) Param Effi.	Accu Δ (FedReFT+(R 8) vs. others)
FLoRA	32	243.15	78.83	22.09×	−2.61%	88.42×	−3.17%
FedIT	32	48.63	75.74	4.42×	+0.48%	17.68×	−0.08%
FFA-LoRA	32	24.31	71.11	2.21×	+5.11%	8.84×	+4.55%
Fed-SB	120	2.83	75.66	0.26×	+0.56%	1.03×	0.00%
FedReFT+	32	11.01	76.22	—	—	4.00×	+0.56%
FedReFT+	8	2.75	75.66	0.25×	−0.56%	—	—

Question 2: Table 4 displays the performance of FedReFT+ across different models with two setups. However, it is unclear what the baseline performance is under these settings.

Answer: We did not find any reference work on arithmetic reasoning in the Distinct Task and Mixed Task experiments setup. We found some baseline work on the arithmetic reasoning task over the GSM8K dataset, and we compare FedReFT+ in Table 6 for the GSM8K Dataset. FedReFT+ achieves 7.25× and 3.63× parameter efficiency, along with +3.05% and +3.36% accuracy improvements over FedSA-LoRA and FFA-LoRA, respectively.

Question 3: Table 5 only presents performance on four GLUE subtasks. A more comprehensive comparison across all subtasks would provide a clearer analysis.

Answer: We have taken the performance results of all baseline methods from (Guo et al., 2024). Therefore, we did experiments on these six GLUE subtasks to compare with the well-established baselines. These results across six tasks demonstrate significant performance gains, achieving 27.17× to 49.43× parameter efficiency compared to the SOTA baselines.

Method	RANK	Param (M)	FedReFT+ Param Effi.	MNLI-m	MNLI-mm	SST-2	QNLI	QQP	RTE	Avg
FFA-LoRA	8	1.44	27.17×	88.83	88.27	94.95	91.52	86.71	86.08	89.39
FedDPA-LoRA	8	2.62	49.43×	88.99	88.43	95.50	90.74	85.73	87.44	89.47
FedSA-LoRA	8	1.83	34.53×	90.18	88.88	96.00	92.13	87.48	87.93	90.43
FedReFT+	1	0.053	-	88.86	89.61	95.17	94.52	86.57	87.80	**90.46**

Question 4: Figure 1 is unclear, presenting both the count and percentage of trainable parameters. Since these two metrics are essentially the same, a performance comparison relative to the number of trainable parameters would be more intuitive.

Answer: We redesign Figure 1, which Illustration of the relationship between the average accuracy (in $%$) and trainable parameter (in $%$) for various federated PEFT methods on Commonsense, Arithmetic, and

GLUE benchmarks using LLaMA-3.2B, LLaMA-3 8B, and RoBERTa-large models, respectively. The figure can be found in this anonymous URL

Comments, Suggestions, And Typos:

- Section 2.1 compares various intervention parameter sharing strategies, but FedReFT+ ultimately selects Full Intervention Sharing without any special design. Therefore, this discussion could be moved to the Appendix. In contrast, the ablation study on the Aggregation method is important and should be included in the main body of the paper. Answer: We will correct the typos in the next version. We are considering moving Section 2.1 to the Appendix to improve the flow of the main content. Additionally, we will incorporate the ablation study on the aggregation method into the main paper to highlight its significance.

- Many tables and figures are not located on the same pages where they are referenced, which makes the paper hard to follow. Answer: We sincerely acknowledge this issue and will address it in the next version of the paper.