Thank you to all reviewers for the careful reading, constructive feedback, and positive assessment of the paper. We are glad that the main contribution was recognized as a principled new post-training compression paradigm for LLMs: replacing single-subspace low-rank factorization with activation-aware sparse dictionary learning, yielding a more expressive union-of-subspaces representation. We also appreciate that the empirical improvements over strong SVD-based baselines were found convincing overall.

At the same time, the reviewers raised several valuable points that helped us improve the paper. The main issues concerned compression-time overhead, practical deployment efficiency, reproducibility details, storage accounting, broader evaluation, comparison to semi-structured sparse baselines, positioning relative to quantization-based methods, and the use of uniform rather than dynamic budget allocation. We believe we have addressed all of these concerns in our responses.

More specifically:

- **Reviewer xicA** asked for clearer implementation details, more transparent storage accounting, stronger discussion of runtime efficiency, and a structured sparse baseline. In response, we clarified the solver configuration and reproducibility details, prepared an explicit storage breakdown, added end-to-end throughput results, and included a comparison against Wanda 2:4.

- **Reviewer X2Ep** raised concerns about offline compression cost, the gap between asymptotic and realized efficiency, and uniform budget allocation. We clarified these points, discussed standard acceleration routes and sparse-kernel relevance, and explained why dynamic allocation is complementary rather than contradictory to the main contribution. Importantly, this reviewer later confirmed that they had no further concerns.

- **Reviewer 53eL** asked about runtime, evaluation breadth, and the relation to quantization-based compression. In response, we added a runtime breakdown, highlighted broader reasoning-oriented benchmarks already included in the appendix, and clarified that CoSpaDi and PTQ operate on different but complementary compression axes.

In the final manuscript, we plan to:

- consolidate implementation and reproducibility details in one clearly visible place;

- clarify storage conventions with an explicit breakdown of dictionary, coefficient, and mask/index costs;

- strengthen the discussion of practical deployment efficiency and runtime limitations;

- add the extra throughput and semi-structured baseline comparisons;

- promote broader benchmark coverage from the appendix into the main paper;

- improve the discussion of the relation between CoSpaDi and quantization-based post-training compression.

Overall, we believe the reviewer feedback has substantially improved the clarity and positioning of the work, and we are grateful for it.

If we have overlooked any point, or if the reviewers would still like to see any additional clarification, we would be very happy to incorporate further suggestions.