

000 QUANTIZED CACHE-TO-CACHE: COMMUNICATION- 001 BUDGETED KV TRANSFER FOR HETEROGENEOUS 002 LLMS

006 **Anonymous authors**

007 Paper under double-blind review

011 ABSTRACT

013 We study communication-efficient transfer between heterogeneous large language
014 models (LLMs) by quantizing Cache-to-Cache (C2C) KV-cache transfer. Our goal
015 is to reduce bandwidth and memory while preserving accuracy. We present post-
016 training quantization (INT8/INT4), cache-length reduction, and accuracy-versus-
017 bytes curves for a heterogeneous model pair. Empirically, quantization is nearly
018 lossless, while cache-length pruning reveals a strong front/back asymmetry that is
019 critical for budgeted transfer. We release a reproducible evaluation pipeline and
020 analysis scripts, and we outline a main-conference path toward sparse, projector-
021 aware token selection and mixed precision.

023 1 INTRODUCTION

025 Large language models (LLMs) often communicate via text, which is slow and lossy. Cache-to-
026 Cache (C2C) communicates via KV-cache projection and fusion, but does not address precision or
027 bandwidth constraints. We ask: *How low can KV precision go before accuracy collapses, and can*
028 *we recover performance under tight communication budgets?*

030 Contributions.

- 032 • We introduce a precision-aware C2C evaluation pipeline and quantify INT8/INT4 PTQ effects on
C2C accuracy.
- 034 • We study cache-length reduction as a second budget axis and show that back-pruning consistently
outperforms front-pruning.
- 036 • We report accuracy vs. communication-budget curves that jointly compare precision and cache
length.
- 038 • We provide a reproducible benchmarking setup and analysis scripts to support extensions to QAT,
mixed precision, heterogeneity, and selective transfer.

041 2 BACKGROUND AND MOTIVATION

044 C2C projects sharer KV caches into receiver space and fuses them with learned gates, preserving rich
045 semantics compared to text relay. However, KV caches are large: they scale with sequence length,
046 KV heads, and head dimension. Quantization and cache-length reduction can shrink the communica-
047 tion footprint while retaining accuracy. This work reframes C2C through a communication-budget
048 lens.

049 3 RELATED WORK

052 **C2C.** Cache-to-Cache (C2C) enables direct semantic communication by projecting and fusing a
053 sharer model’s KV cache into a receiver’s KV cache with learnable gates, avoiding intermediate text
generation (Fu et al., 2025).

KV communication across agents. KVComm aligns KV caches across diverging prefixes using training-free offset correction with online anchors (Ye et al., 2025). Q-KVComm adds adaptive layer-wise quantization, hybrid information extraction, and heterogeneous calibration for compressed KV transfer (Kriuk & Ng, 2025). These works focus on multi-agent cache reuse/compression; our work studies quantization and cache-length pruning within the C2C projector+fuser pipeline.

Latent collaboration and cache alignment. KV cache alignment learns a shared latent space with adapters to align KV caches across models (Dery et al., 2026). LatentMAS enables latent-space collaboration with shared working memory without extra training (Zou et al., 2025). Our approach stays within C2C’s KV fusion but emphasizes communication budgets and precision/length trade-offs.

Token selection and KV compression. Token-level KV selection and value-norm importance improve long-context inference for a single model (ZipCache, TokenSelect, VATP) (Anonymous, 2024b; 2025; 2024a). We adopt the budget perspective for C2C rather than single-model KV compression.

4 METHOD

4.1 C2C RECAP

Let the sharer model produce KV caches (K_ℓ^S, V_ℓ^S) and the receiver produce (K_ℓ^R, V_ℓ^R) at layer ℓ . C2C projects sharer KV into receiver space via Π_ℓ^K, Π_ℓ^V and fuses them through a learnable gate:

$$(K_\ell^{R'}, V_\ell^{R'}) = \mathcal{F}_\ell(K_\ell^R, V_\ell^R, \Pi_\ell^K(K_\ell^S), \Pi_\ell^V(V_\ell^S)).$$

This avoids intermediate text and transfers richer internal semantics.

4.2 POST-TRAINING QUANTIZATION (PTQ)

We quantize the KV caches using INT8 or INT4/NF4 with per-head scaling. We evaluate accuracy and latency under fixed precision budgets. Our current implementation uses fake-quant (quantize then dequantize) to model quantization noise without bit-packing.

4.3 CACHE-LENGTH REDUCTION

We prune KV tokens using a fixed ratio (e.g., 50%, 25%, 10%), reducing transmitted bytes further. We evaluate front-pruning and back-pruning to diagnose which instruction tokens are most valuable for cross-model transfer.

4.4 SELECTIVE AND COMPRESSED CACHE TRANSFER (SPARSEC2C)

As a main-conference extension, we select a sparse subset of token positions to transfer and fuse. Let $I \subset \{1, \dots, T\}$ be selected tokens and S_I the gather operator. We fuse only selected tokens and scatter updates back:

$$\begin{aligned} (\tilde{K}_\ell^R, \tilde{V}_\ell^R) &= S_I^\top(K_\ell^R, V_\ell^R), & (\tilde{K}_\ell^S, \tilde{V}_\ell^S) &= S_I^\top(K_\ell^S, V_\ell^S) \\ (\tilde{K}_\ell^{R'}, \tilde{V}_\ell^{R'}) &= \mathcal{F}_\ell(\tilde{K}_\ell^R, \tilde{V}_\ell^R, \Pi_\ell^K(\tilde{K}_\ell^S), \Pi_\ell^V(\tilde{V}_\ell^S)). \end{aligned}$$

We then scatter the update to the full cache. We use projector-aware token scoring by computing value norms in receiver space (`proj_vnorm_topk`), tying selection to the cross-model mapping.

4.5 COMMUNICATION-BUDGET CURVES

We report accuracy as a function of transmitted bytes, enabling fair comparison under equal communication constraints. For a sequence of length T , the approximate bytes are

$$\text{bytes} \approx T \cdot p \cdot 2 \cdot L \cdot H_{kv} \cdot d_h \cdot b/8,$$

where p is the retained cache proportion, L the number of layers, H_{kv} KV heads, d_h head dim, and b bits per element. We use this accounting for consistent budget curves.

108 5 EXPERIMENTS
109110 5.1 SETUP
111

112 We evaluate on OpenBookQA and ARC-C with a Qwen3-0.6B receiver and Qwen2.5-0.5B sharer.
 113 We follow the C2C eval protocol: temperature 0, max_new_tokens 64, no CoT, unified chat template.
 114 All models are frozen; only the projector is trained when QAT is enabled. The OpenBookQA test
 115 split has 500 samples and ARC-C has 1172 samples.

116 5.2 MAIN RESULTS
117

118 All results below are full runs. PTQ is effectively lossless relative to FP16, and cache pruning shows
 119 a strong front/back asymmetry.

122 Table 1: Baseline vs. PTQ (full-cache, %).
123

| Setting | OpenBookQA | ARC-C |
|---------------|------------|-------|
| FP16 baseline | 52.8 | 55.1 |
| INT8 PTQ | 52.8 | 55.0 |
| INT4 PTQ | 52.6 | 55.4 |

128 129 Table 2: OpenBookQA accuracy (%), 500 samples) for cache-length pruning (INT8).
130

| Order mode | 75% | 50% | 25% | 10% |
|------------|------|------|------|------|
| Front | 44.6 | 43.0 | 38.8 | 38.6 |
| Back | 52.2 | 52.0 | 50.8 | 49.2 |

135 136 Table 3: ARC-C accuracy (%), 1150 samples) for cache-length pruning (INT8).
137

| Order mode | 75% | 50% | 25% | 10% |
|------------|------|------|------|------|
| Front | 40.2 | 46.3 | 38.3 | 40.7 |
| Back | 55.7 | 57.2 | 56.2 | 53.7 |

143 5.3 COMMUNICATION-BUDGET CURVE
144

145 Figure 1 and Figure 2 report accuracy versus effective transmitted bytes. Each point is annotated
 146 with the retained cache proportion. These curves provide a single, comparable view across precision
 147 (FP16/INT8/INT4) and cache-length reduction.

148 5.4 ORDER-MODE ABLATION
149

150 Across all cache lengths, **back-pruning** (keeping later instruction tokens) consistently outperforms
 151 **front-pruning**. At 50% cache length, for example, back-pruning retains near-baseline accuracy
 152 while front-pruning degrades sharply. This suggests late instruction tokens carry higher utility for
 153 cross-model KV fusion, a useful design signal for future selective transfer methods.

154 5.5 MAIN-CONFERENCE EXTENSIONS (PRELIMINARY)
155

156 We report early results for two main-conference extensions. Mixed precision (INT8 with FP16 in
 157 the last layers) remains near baseline. An alignment-only ablation (same model pair, alignment
 158 enabled) reduces accuracy, suggesting alignment should be reserved for heterogeneous pairs. For
 159 SparseC2C (token selection), INT4 with vnorm selection remains strong: at p=0.5 we observe 52.0
 160 on OpenBookQA and 56.4 on ARC-C, while front pruning degrades sharply. Full INT8 grids and
 161 heterogeneity runs will be reported in the main-conference track.

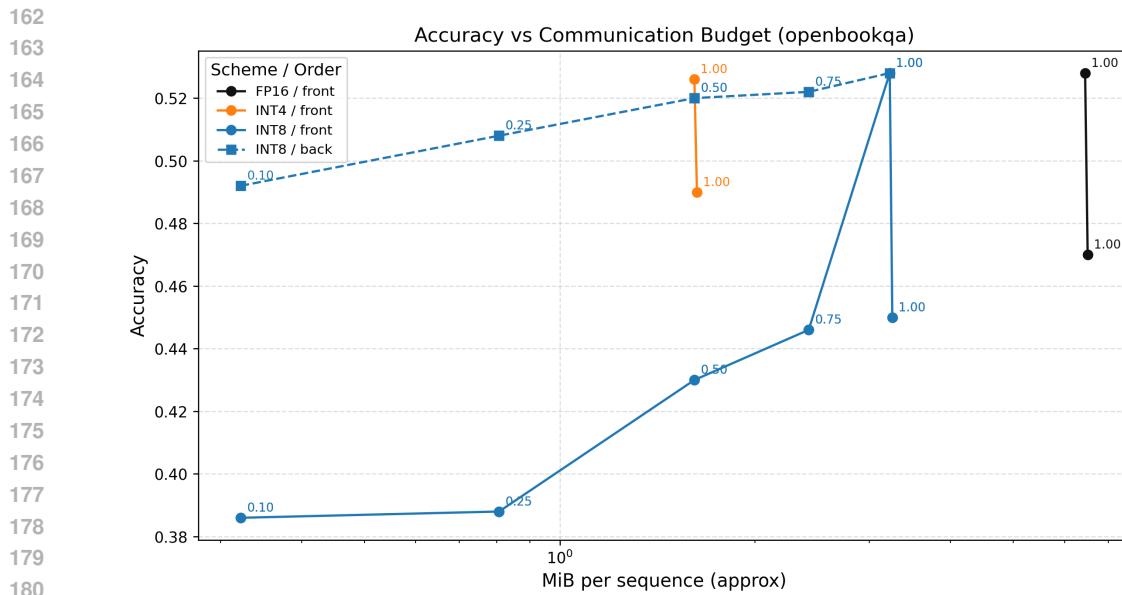


Figure 1: Accuracy vs. communication budget (OpenBookQA).

Figure 2: Accuracy vs. communication budget (ARC-C).

6 DISCUSSION

Quantized C2C provides large bandwidth reductions with limited accuracy drop. Cache pruning further improves the tradeoff, suggesting a practical path to deployable multi-LLM communication. A main-conference path includes QAT recovery, mixed-precision schedules, heterogeneous model pairs, and SparseC2C token selection.

216
217
218
219
220
221
222

Table 4: Preliminary extension results (% accuracy).

| Setting | OpenBookQA | ARC-C |
|--------------------------------------|------------|-------|
| Mixed precision (INT8 + last-4 FP16) | 52.8 | 55.3 |
| Alignment ablation (same pair) | 46.8 | 49.6 |

7 LIMITATIONS

Our results currently focus on a single model pair and two datasets. We do not yet report end-to-end latency or FLOP measurements for the fuser, and SparseC2C remains an ongoing extension. These limitations will be addressed in the main-conference track.

8 BROADER IMPACT

Communication-efficient multi-LLM systems can reduce compute and latency, but they may also enable higher-throughput deployment of models. We emphasize reproducible evaluation, careful reporting of accuracy/latency tradeoffs, and responsible deployment in sensitive domains.

9 CONCLUSION

We introduce precision-aware C2C and report accuracy vs. bytes curves. This establishes a communication-budget perspective for cross-model KV transfer and opens the door to low-latency, low-bandwidth agent collaboration.

ACKNOWLEDGMENTS

Placeholder.

REFERENCES

Anonymous. Attention score is not all you need for token importance indicator in kv cache reduction: Value also matters. 2024a.

Anonymous. Zipcache: Accurate and efficient kv cache compression for llm inference. 2024b.

Anonymous. Tokenselect: Efficient long-context inference with token-level kv cache selection. 2025.

Lucio M. Dery, Zohar Yahav, Henry Prior, Qixuan Feng, Jiajun Shen, and Arthur Szlam. Latent space communication via k-v cache alignment. 2026.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. 2025.

Boris Kriuk and Logic Ng. Q-kvcomm: Efficient multi-agent communication via adaptive kv cache compression. 2025.

Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, and Yiran Chen. Kvcomm: Online cross-context kv-cache communication for efficient llm-based multi-agent systems. 2025.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. 2025.

268
269