Novelty Assessment Report

Paper: LS-Merge: Merging Language Models in Latent Space
PDF URL: https://openreview.net/pdf?id=VSDV0SWwOC
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Model merging in weight space is an efficient way to reuse pretrained models, but existing methods typically assume matching architectures or sizes, making heterogeneous merges brittle or infeasible. We address this limitation by encoding model weights into a smooth latent space, enabling cross-architecture operations, and performing the merge in the latent space before decoding back to weights. This approach faces two major challenges. First, LLMs contain billions of parameters, which makes latent encoding computationally demanding. Second, using high compression ratios often hinders the encoder's ability to generalize to unseen weights. We tackle these issues with a transformer-based variational autoencoder (VAE) trained in a two-stage compression curriculum with structured layer-aware chunking: the model first learns a high-capacity latent representation and then distills to a compact code, improving both stability and out-of-distribution generalization. To align heterogeneous models, we introduce a dimensionality-matching projection that allows interpolation between models of different sizes. Empirically, latent-space interpolation is consistently more robust than direct weight-space averaging and yields stronger downstream performance when merging models of different sizes. Together, these components provide a scalable, architecture-agnostic recipe for model merging.

Core Task Landscape

This paper addresses: **Merging Language Models with Heterogeneous Architectures**

A total of **50 papers** were analyzed and organized into a taxonomy with **22 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Parameter-Space Merging and Alignment Techniques**
- **Knowledge Transfer and Ensemble Collaboration**
- **Mixture-of-Experts Architectures for Heterogeneous Models**
- **Multimodal and Cross-Domain Model Integration**
- **Domain-Specific Merging Applications**
- **Optimization and Evaluation Frameworks**
- **Auxiliary Techniques and Supporting Methods**
- **Peripheral and Tangentially Related Work**

Complete Taxonomy Tree

- Merging Language Models with Heterogeneous Architectures Survey Taxonomy
- Parameter-Space Merging and Alignment Techniques
  - Latent-Space and Embedding-Based Merging ★ (3 papers)
  - [0] LS-Merge: Merging Language Models in Latent Space (Anon et al., 2026) View paper
  - [1] Mergenet: Knowledge migration across heterogeneous models, tasks, and modalities (Zhan Tianyu, 2025) View paper
  - [2] Knowledge fusion of large language models (Wan, 2024) View paper
  - Weight-Space Interpolation and Coefficient Optimization (3 papers)
  - [6] AdaMMS: Model Merging for Heterogeneous Multimodal Large Language Models with Unsupervised Coefficient Optimization (Yiyang Du, 2025) View paper
  - [24] StatsMerging: Statistics-Guided Model Merging via Task-Specific Teacher Distillation (Ranjith Merugu, 2025) View paper
  - [25] Twin-Merging: Dynamic Integration of Modular Expertise in Model Merging (Lu Zhenyi, 2024) View paper
  - Layer-Level Integration and Permutation (2 papers)
  - [49] Model Assembly Learning with Heterogeneous Layer Weight Merging (Zhang, 2025) View paper
  - [50] Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models (Bandarkar, 2024) View paper
- Knowledge Transfer and Ensemble Collaboration
  - Knowledge Distillation and Amalgamation (2 papers)
  - [3] Amalgamating Multi-Task Models with Heterogeneous Architectures (Jidapa Thadajarassiri, 2024) View paper
  - [32] Fusechat: Knowledge fusion of chat models (Wan, 2025) View paper
  - Multi-Model Ensemble Inference (3 papers)
  - [5] Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration (Xiaocheng Feng, 2024) View paper
  - [9] Mixture-of-Agents Enhances Large Language Model Capabilities (Wang Junlin, 2024) View paper
  - [34] Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models (Lu Jinliang, 2024) View paper
- Mixture-of-Experts Architectures for Heterogeneous Models
  - MoE Construction and Expert Routing (2 papers)

    - [8] MergeME: Model Merging Techniques for Homogeneous and Heterogeneous MoEs (Zhou, 2025) View paper
    - [43] Mixture of experts in large language models (Zhang Danyang, 2025) View paper
    - MoE Deployment and Resource Optimization (1 papers)
    - [42] CoMoE: Collaborative Optimization of Expert Aggregation and Offloading for MoE-based LLMs at Edge (Li Muqing, 2025) View paper
- Multimodal and Cross-Domain Model Integration
    - Vision-Language Model Fusion (5 papers)
    - [4] DeepMLF: Multimodal language model with learnable tokens for deep fusion in sentiment analysis (Georgiou, 2025) View paper
    - [22] GeoLangBind: Unifying Earth Observation with Agglomerative Vision-Language Foundation Models (Xiong, 2025) View paper
    - [30] SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation (Li Wei-Hua, 2024) View paper
    - [40] SAM2CLIP2SAM: Vision Language Model for Segmentation of 3D CT Scans for Covid-19 Detection (Kollias, 2024) View paper
    - [47] Dual-Branch Knowledge Enhancement Network with Vision-Language Model for Human-Object Interaction Detection (Guangpu Zhou, 2024) View paper
    - Audio-Language Integration (2 papers)
    - [10] AudioPaLM: A Large Language Model That Can Speak and Listen (Rubenstein, 2023) View paper
    - [28] LaMoSC: Large Language Model-Driven Semantic Communication System for Visual Transmission (Yaru Zhao, 2024) View paper
    - Sign Language Translation (1 papers)
    - [20] Sign Language Translation using Hybrid Temporal Convolutional Network with TTS Fusion (K Ananthajothi, 2025) View paper
    - Text-Based Multimodal Fusion (3 papers)
    - [19] Effective and Explainable Molecular Property Prediction by Chain-of-Thought Enabled Large Language Models and Multi-Modal Molecular Information Fusion (Chang Jin, 2025) View paper
    - [29] Ethereum Fraud Detection via Joint Transaction Language Model and Graph Representation Learning (Sun Jian-guo, 2024) View paper
    - [33] Multi-scale feature fusion and graph neural network integration for text classification with large language models (Xiangchen Song, 2025) View paper
- Domain-Specific Merging Applications
    - Scientific and Technical Domain Adaptation (1 papers)
    - [7] Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities (Wei Lu, 2024) View paper
    - Cross-Lingual Model Merging (2 papers)
    - [36] Extend Model Merging from Fine-Tuned to Pre-Trained Large Language Models via Weight Disentanglement (Yu Le, 2024) View paper
    - [46] Mitigating Catastrophic Forgetting in Language Transfer via Model Merging (Raychev Veselin, 2024) View paper
    - Specialized Application Domains (4 papers)
    - [14] Text2BIM: Generating Building Models Using a Large Language Model-Based Multiagent Framework (Du Chang-yu, 2026) View paper
    - [17] Hierarchical Large Language Models in Cloud-Edge-End Architecture for Heterogeneous Robot Cluster Control (Zhirong Luan, 2023) View paper
    - [26] (In) forming the new building envelope: A pedagogical study in generative design with precedents and multimodal large language models (Pedro Veloso, 2025) View paper
    - [39] MambaLLM: Integrating Macro-Index and Micro-Stock Data for Enhanced Stock Price Prediction (Jin Yan, 2025) View paper
- Optimization and Evaluation Frameworks
    - Evolutionary and Automated Merging Optimization (1 papers)
    - [16] Evolutionary optimization of model merging recipes (Takuya Akiba, 2024) View paper
    - Comprehensive Survey Studies (2 papers)
    - [15] Democratizing AI through model fusion: A comprehensive review and future directions (Qi Zhou, 2025) View paper
    - [23] Deep model fusion: A survey (Li Weishi, 2023) View paper
- Auxiliary Techniques and Supporting Methods
    - Semantic Alignment and Representation Learning (1 papers)
    - [37] Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts? (Crystina Zhang, 2024) View paper
    - Dynamic Model Editing and Adaptation (1 papers)
    - [41] DAFNet: Dynamic Auxiliary Fusion for Sequential Model Editing in Large Language Models (Zhang, 2024) View paper
    - Guardrails and Safety Mechanisms (1 papers)
    - [35] RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content (Yuan, 2024) View paper
- Peripheral and Tangentially Related Work
    - Cross-Lingual Transfer and Typology Analysis (1 papers)
    - [18] Untangling the Influence of Typology, Data and Model Architecture on Ranking Transfer Languages for Cross-Lingual POS Tagging (Enora Rice, 2025) View paper
    - Specialized Domain Applications with Minimal Merging (5 papers)
    - [12] LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers (Theo Olausson, 2023) View paper
    - [13] Spatiotemporal Pretrained Large Language Model for Forecasting With Missing Values (Le Fang, 2025) View paper
    - [31] Autism Spectrum Disorder Diagnosis Through Natural Language Processing and ConvCapsule Fusion Network (Jagadesh Balasubramani, 2025) View paper
    - [44] Artificial intelligence applications in education: Natural language processing in detecting misconceptions (Yunus KÃ¶kver, 2024) View paper
    - [48] GCN-LSTM: multi-label educational emotion prediction based on graph convolutional network and long and short term memory network fusion label correlation in â□¦ (Z Liu, 2024) View paper
    - Architectural References and Tangential Mentions (5 papers)
    - [11] Decision-Making Large Language Model for Wireless Communication: A Comprehensive Survey on Key Techniques (Ning Yang, 2025) View paper

- [21] LLM-WFIN: A Fine-Grained Large Language Model (LLM)-Oriented Website Fingerprinting Attack via Fusing Interrupt Trace and Network Traffic (Jiajia Jiao, 2025) View paper
- [27] A survey on intelligent network operations and performance optimization based on large language models (Sifan Long, 2025) View paper
- [38] Command A: An Enterprise-Ready Large Language Model (- -, 2025) View paper
- [45] Merging and processing heterogeneous models (Dissaux, 2016) View paper

Narrative

Core task: merging language models with heterogeneous architectures. The field has evolved to address the challenge of combining models that differ in structure, training objectives, or domain specialization. The taxonomy reveals several complementary strategies: Parameter-Space Merging and Alignment Techniques focus on direct weight manipulation and embedding-based fusion methods such as those explored in LS-Merge[0] and Mergenet[1]; Knowledge Transfer and Ensemble Collaboration emphasize collaborative inference and distillation approaches like Ensemble Heterogeneous LLMs[5] and Mixture-of-Agents[9]; Mixture-of-Experts Architectures provide modular frameworks for routing across heterogeneous components; while Multimodal and Cross-Domain Model Integration tackles the broader challenge of fusing models trained on different modalities or tasks. Domain-Specific Merging Applications and Optimization and Evaluation Frameworks address practical deployment and benchmarking concerns, with supporting methods in auxiliary techniques rounding out the landscape.

A particularly active line of work centers on latent-space and embedding-based merging, where models are aligned through learned transformations rather than naive parameter averaging. LS-Merge[0] exemplifies this direction by operating in a shared latent space to bridge architectural differences, positioning itself alongside Mergenet[1] which also emphasizes learned alignment mechanisms, and Knowledge Fusion LLMs[2] which explores fusion at the representation level. These approaches contrast with ensemble methods like Mixture-of-Agents[9] that preserve model independence during inference, and with mixture-of-experts frameworks such as CoMoE[42] that introduce explicit gating. A recurring theme across branches is the trade-off between integration depth—whether to merge parameters directly, align intermediate representations, or coordinate outputs—and the preservation of specialized capabilities. Open questions include how to efficiently search the space of possible merges, as addressed by Evolutionary Model Merging[16], and how to evaluate merged models across diverse benchmarks without retraining from scratch.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Mergenet: Knowledge migration across heterogeneous models, tasks, and modalities

Authors: Zhan Tianyu, Kunxi Li, Tianyu Zhan, Zhang Shengyu, Shengyu Zhang, et al. (16 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

In this study, we focus on heterogeneous knowledge transfer across entirely different model architectures, tasks, and modalities. Existing knowledge transfer methods (e.g., backbone sharing, knowledge distillation) often hinge on shared elements within model structures or task-specific features/labels, limiting transfers to complex model types or tasks. To overcome these challenges, we present MergeNet, which learns to bridge the gap of parameter spaces of heterogeneous models, facilitating the ...

Relationship Analysis

Both papers belong to the Latent-Space and Embedding-Based Merging category, encoding model parameters into latent representations to enable cross-architecture merging. They overlap in addressing heterogeneous model merging by projecting weights into shared latent spaces, but differ fundamentally in approach: LS-Merge uses a transformer-based VAE with optimal transport alignment to merge entire pretrained LLMs in a unified latent space, while MergeNet employs low-rank decomposition and attention-based parameter adapters to transfer knowledge between specific layers of heterogeneous models during training.

2. Knowledge fusion of large language models

Authors: Wan, Fanqi, Fanqi Wan, Huang Xinting, Xinting Huang, et al. (17 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

While training large language models (LLMs) from scratch can generate models with distinct functionalities and strengths, it comes at significant costs and may result in redundant capabilities. Alternatively, a cost-effective and compelling approach is to merge existing pre-trained LLMs into a more potent model. However, due to the varying architectures of these LLMs, directly blending their weights is impractical. In this paper, we introduce the notion of knowledge fusion for LLMs, aimed at com...

Relationship Analysis

Both papers belong to the Latent-Space and Embedding-Based Merging category, using learned representations to enable model merging operations. They overlap in addressing cross-architecture merging challenges by projecting model parameters into shared latent spaces. However, the original paper (LS-Merge) encodes raw weight tensors into latent codes using a transformer-based VAE and performs interpolation in this compressed latent space, while the candidate paper (Knowledge Fusion) operates on the probabilistic output distributions of LLMs rather than their weights, fusing these generative distributions to transfer knowledge to a target model through continual training.

Contributions Analysis

Overall novelty summary. The paper proposes LS-Merge, a framework that encodes model weights into a latent space using a transformer-based VAE, enabling cross-architecture merging operations before decoding back to weights. This work resides in the 'Latent-Space and Embedding-Based Merging' leaf, which contains only three papers including the original. This is a relatively sparse research direction within the broader taxonomy of 50 papers across 22 leaf nodes, suggesting that latent-space encoding approaches for heterogeneous model merging remain an emerging area compared to more established techniques like direct weight interpolation or ensemble methods.

The taxonomy reveals that this work sits within 'Parameter-Space Merging and Alignment Techniques', adjacent to 'Weight-Space Interpolation and Coefficient Optimization' (3 papers) and 'Layer-Level Integration and Permutation' (2 papers). These neighboring leaves focus on direct parameter manipulation without latent encoding, highlighting a methodological divergence. The broader taxonomy also includes 'Knowledge Transfer and Ensemble Collaboration' (5 papers) and 'Mixture-of-Experts Architectures' (3 papers), which preserve model independence rather than merging parameters. The scope notes clarify that latent-space methods explicitly exclude direct weight averaging and ensemble approaches, positioning this work as a distinct strategy for achieving heterogeneous integration through learned representations.

Among 29 candidates examined, the contribution-level analysis reveals mixed novelty signals. The core LS-Merge framework (Contribution 1) examined 10 candidates with 1 appearing to provide overlapping prior work. The dimensionality-matching projection and optimal transport alignment (Contribution 2) examined 9 candidates with 2 potentially refuting papers. The two-stage compression curriculum with layer-aware chunking (Contribution 3) examined 10 candidates with none clearly refuting it, suggesting this training

strategy may be more novel within the limited search scope. These statistics indicate that while the overall latent-space merging concept has some precedent, specific technical components—particularly the compression curriculum—appear less explored in the examined literature.

Based on the limited search of 29 semantically similar papers, the work appears to occupy a relatively sparse research direction with modest prior overlap. The taxonomy structure confirms that latent-space encoding for heterogeneous merging is less crowded than direct weight-space methods or ensemble approaches. However, the analysis does not cover exhaustive literature search or systematic review of all related compression and alignment techniques, leaving open the possibility of additional relevant work outside the top-K semantic matches examined.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: LS-Merge framework for merging LLMs in latent space

Description: The authors introduce LS-Merge, a framework that encodes model weights into a smooth latent space using a transformer-based variational autoencoder, performs merging operations in this latent space, and decodes back to weights. This approach enables both homogeneous and heterogeneous model merging without requiring architectural alignment.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. EvoEdit: Lifelong Free-Text Knowledge Editing through Latent Perturbation Augmentation and Knowledge-driven Parameter Fusion

URL: View paper

Brief Assessment

EvoEdit[69] addresses lifelong knowledge editing in LLMs using latent perturbation augmentation, not model merging. The candidate focuses on updating factual knowledge over time, while the original contribution concerns combining multiple pretrained models by encoding weights into latent space.

---

2. Latent feature transformation for emergent task performance in large language models

URL: View paper

Brief Assessment

Latent Feature Transformation[66] focuses on emergent task performance through latent feature transformation within pre-trained models, not on merging multiple models' weights in latent space as LS-Merge does.

---

3. Latent syntax weaving in large language model representations: A novel mechanism for self-referential consistency in neural architectures

URL: View paper

Brief Assessment

Latent Syntax Weaving[63] focuses on self-referential consistency mechanisms in neural architectures through latent syntax patterns. The candidate's fragmentary mentions of 'latent' relate to attention pooling and topological properties, not to weight-space encoding or model merging frameworks.

---

4. SeMe: Training-Free Language Model Merging via Semantic Alignment

URL: View paper

Prior Art Analysis

SeMe[62] demonstrates that merging language models through latent semantic alignment was proposed prior to the original paper's LS-Merge framework. Both approaches share the core concept of performing model merging operations in a learned latent representation space rather than directly in weight space. SeMe[62] explicitly describes a 'data-free, and training-free approach that leverages latent semantic alignment to merge lms at a fine-grained, layer-wise level,' which directly overlaps with the original paper's claim of introducing latent-space merging. The fundamental innovation of moving from weight-space to latent-space merging, which the original paper presents as novel, appears to have been established by SeMe[62].

Evidence

Evidence 1 - **Rationale**: Both papers claim to introduce novel frameworks for merging language models in latent space. SeMe[62] explicitly describes merging through 'latent semantic alignment' at a 'layer-wise level,' which directly corresponds to the original paper's claim of shifting merging 'from the raw weight space to a learned latent space.' - **Original**: we propose ls-merge, a novel framework that fundamentally shifts the merging process from the raw weight space to a learned latent space. - **Candidate**: we introduce seme (semantic-based merging), a novel, data-free, and training-free approach that leverages latent semantic alignment to merge lms at a fine-grained, layer-wise level.

Evidence 2 - **Rationale**: SeMe[62] positions itself against existing weight-space merging techniques and proposes latent-space merging as an alternative, establishing that this approach existed prior to the original paper's submission. - **Original**: merging llms in latent space.we propose ls-merge, a novel latent-space merging methodology that enables merging llms in their weights latent space. - **Candidate**: existing model merging techniques, such as parameter averaging and task-guided fusion, often rely on data-dependent computations or fail to preserve internal knowledge, limiting their robustness and scalability.

Evidence 3 - **Rationale**: SeMe[62] demonstrates merging 'across diverse architectures,' which indicates support for heterogeneous merging similar to what the original paper claims as a novel capability of their latent-space approach. - **Original**: this paradigm enables both homogeneous and heterogeneous merging by design - **Candidate**: through extensive experiments across diverse architectures and tasks, we demonstrate that seme outperforms existing methods in both performance and efficiency while eliminating reliance on external data.

---

5. EnergyMogen: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space

URL: View paper

Brief Assessment

EnergyMogen[65] focuses on human motion generation using energy-based diffusion models in latent space for composing semantic concepts. This is fundamentally different from LS-Merge's weight-space encoding and merging of language models.

---

6. Emergent semantic entanglement in large language models: Non-sequential contextual weaving through stochastic syntagmatic bridges

URL: View paper

Brief Assessment

Semantic Entanglement LLMs[61] does not address model merging or weight-space operations. The candidate focuses on semantic relationships and contextual weaving in language models, which is unrelated to the LS-Merge framework's approach of encoding model weights into latent space for merging purposes.

---

7. AlignMerge - Alignment-Preserving Large Language Model Merging via Fisher-Guided Geometric Constraints

URL: View paper

Brief Assessment

AlignMerge[70] operates on weight-space merging with Fisher-weighted geometry constraints around an aligned anchor, not on encoding weights into a learned latent space via VAEs. The candidate focuses on preserving alignment during standard weight-space operations, while the original proposes a fundamentally different paradigm of VAE-based latent encoding and decoding.

---

8. Hierarchical Contextual Manifold Alignment for Structuring Latent Representations in Large Language Models

URL: View paper

Brief Assessment

Hierarchical Contextual Manifold[68] focuses on restructuring token embeddings within a single model's latent representation space without modifying weights, while LS-Merge encodes entire model weight tensors into latent space for cross-model merging operations. These are fundamentally different technical approaches addressing distinct problems.

---

9. Flows and Diffusions on the Neural Manifold

URL: View paper

Brief Assessment

Neural Manifold Flows[64] focuses on generating neural network weights through diffusion and flow-based models for trajectory inference and optimization dynamics, not on merging pretrained language models. The candidate addresses weight generation and initialization, while the original addresses weight-space model merging.

---

10. Unsupervised Neural Machine Translation with Weight Sharing

URL: View paper

Brief Assessment

Weight Sharing NMT[67] focuses on unsupervised neural machine translation by encoding sentences into a shared latent space for cross-lingual translation, not on merging language model weights. The latent space serves a fundamentally different purpose (translation vs. model merging).

Contribution 2: Dimensionality-matching projection and OT-based alignment for heterogeneous merging

Description: The authors develop a method combining proportional dimensionality mapping with Optimal Transport alignment to enable merging of models with mismatched architectures (different depths or widths). This addresses the geometric incompatibility of latent distributions from heterogeneous models by registering their manifolds before interpolation.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Fusion of Graph Neural Networks via Optimal Transport

URL: View paper

Brief Assessment

Graph Fusion Transport[80] focuses on fusing Graph Convolutional Networks (GCNs) using optimal transport for layer-wise weight alignment, not on addressing dimensionality mismatches between heterogeneous architectures with different depths or widths as in the original paper.

---

2. A Survey on Optimal Transport for Machine Learning: Theory and Applications

URL: View paper

Brief Assessment

Optimal Transport Survey[79] is a survey paper that reviews OT applications across ML domains including model fusion, but does not propose novel methods for merging neural networks with heterogeneous architectures. The original paper's specific contribution of combining proportional dimensionality mapping with OT alignment for heterogeneous model merging is not addressed in this survey.

---

3. Model fusion via optimal transport

URL: View paper

Prior Art Analysis

Model Fusion Transport[72] demonstrates prior work on using optimal transport for aligning and merging neural networks with different architectures. The candidate paper explicitly addresses heterogeneous model fusion where layer widths differ between models, using optimal transport to align neurons before averaging parameters. This directly challenges the novelty claim of the original paper's dimensionality-matching projection combined with OT-based alignment for heterogeneous merging.

Evidence

Evidence 1 - **Rationale**: Both papers address the problem of merging models with different layer widths. The candidate demonstrates that OT-based fusion can handle different sized models by mapping weights from larger to smaller models via transport maps, which is conceptually similar to the original's dimensionality-matching approach. - **Original**: heterogeneous mapping (depth/width mismatch) when two architectures match, layer by layer, in the number of weight chunks, we employ a single v ae to embed all layers into a common d-dimensional latent space. if the per-layer number and chunk counts differ, we instead deploy separate encoders for ea... - **Candidate**: an advantage of our ot-based fusion is that it allows the layer widths to be different for each input model. here, our procedure first identifies which weights of the bigger model should be mapped to the smaller model (via the transport map), and then averages the aligned models (now both of the size ...

Evidence 2 - **Rationale**: Both papers identify the core problem of misaligned latent distributions/neuron correspondences in heterogeneous models and propose optimal transport as the solution for alignment before merging. The candidate explicitly addresses the case where neuron numbers differ across models. - **Original**: optimal transport alignment.while latent encoding standardizes per-layer dimensionality, it does not guarantee that two models' latent representations are geometrically compatible. as shown in appendix c (figure 9a), homogeneous models (e.g., checkpoints from the same pretraining run) exhibit overla... - **Candidate**: motivation. as alluded to earlier in the introduction, the problem with vanilla averaging of parameters is the lack of one-to-one correspondence between the

model parameters. in particular, for a given layer, there is no direct matching between the neurons of the two models. for e.g., this means tha...

Evidence 3 - **Rationale**: The candidate paper's abstract explicitly describes using optimal transport for layer-wise alignment before parameter averaging, which is the core mechanism claimed as novel in the original paper for handling heterogeneous architectures. - **Original**: we introduce a dimensionality-matching projection and ot-based latent alignment that enables interpolation between models of different depths or widths. - **Candidate**: we present a layer-wise model fusion algorithm for neural networks that utilizes optimal transport to (soft-) align neurons across the models before averaging their associated parameters. we show that this can successfully yield "one-shot" knowledge transfer (i.e, without requiring any retraining) b...

---

4. SuperGlue: Learning Feature Matching With Graph Neural Networks

URL: View paper

Brief Assessment

SuperGlue[77] addresses feature matching between image keypoint sets using optimal transport for assignment, not neural network weight merging across heterogeneous architectures. The domains and technical problems are fundamentally different.

---

5. Transformer fusion with optimal transport

URL: View paper

Prior Art Analysis

Transformer Optimal Transport[71] demonstrates prior work on using optimal transport alignment for merging neural networks with different architectures. The candidate paper explicitly addresses heterogeneous fusion of models with different widths using optimal transport theory, presenting a systematic approach that aligns weight matrices before fusion. The paper shows that their OT-based method 'accommodates the fusion of models with different widths, and in turn, different sizes' and provides experimental validation of heterogeneous merging on CIFAR100 where models of different dimensions are successfully fused.

Evidence

Evidence 1 - **Rationale**: Both papers claim to enable merging of models with different widths/sizes. The candidate explicitly states this capability was achieved in 2024, before the original paper's submission. - **Original**: we introduce a dimensionality-matching projection and ot-based latent alignment that enables interpolation between models of different depths or widths. - **Candidate**: additionally, otfusion accommodates the fusion of models with different widths, and in turn, different sizes, which is fundamentally not possible with vf. this is a crucial feature, as such heterogeneous models are available in plenty, to better unleash the potential of existing pre-trained models.

Evidence 2 - **Rationale**: Both papers demonstrate practical implementation of heterogeneous fusion with different model widths, showing successful knowledge transfer. - **Original**: heterogeneous mapping (depth/width mismatch) when two architectures match, layer by layer, in the number of weight chunks, we employ a single v ae to embed all layers into a common d-dimensional latent space. if the per-layer number and chunk counts differ, we instead deploy separate encoders for ea... - **Candidate**: our methodology, as opposed to vf, works out of the box with models having different widths (heterogeneous fusion). we find a consistent absolute increase in test accuracy over the performance of the smaller anchor network, thus implying successful knowledge transfer (tab. 4). these results showcase...

Evidence 3 - **Rationale**: Both papers use optimal transport to ensure geometric compatibility and proper alignment before merging models with different architectures. - **Original**: optimal transport alignment.while latent encoding standardizes per-layer dimensionality, it does not guarantee that two models' latent representations are geometrically compatible. - **Candidate**: we propose to use optimal transport fusion (otfusion) (singh & jaggi, 2020), which at its core, aligns the weight or parameter matrices before fusing them. thus, by virtue of such an alignment, otfusion ensures that the fused model effectively integrates the knowledge and capabilities of the individ...

---

6. Merging embedded topics with optimal transport for online topic modeling on data streams

URL: View paper

Brief Assessment

Embedded Topics Transport[75] applies optimal transport to merge topic embeddings in online topic modeling, not neural network weight merging. The candidate operates on topic distributions in text streams, while the original addresses heterogeneous neural network architecture merging.

---

7. Unsupervised Learning for Optimal Transport plan prediction between unbalanced graphs

URL: View paper

Brief Assessment

The candidate paper (Unbalanced Graph Transport[78]) focuses on learning optimal transport plans between graph structures for graph alignment tasks, not on merging neural network models with different architectures. The technical domains are fundamentally different.

---

8. Graph optimal transport for cross-domain alignment

URL: View paper

Brief Assessment

Graph Cross-Domain Transport[76] focuses on cross-domain alignment between entities in vision-language tasks (e.g., image-text retrieval, VQA) using optimal transport for graph matching. The original paper addresses merging neural network weights with different architectures using dimensionality projection and OT alignment in latent space. These are fundamentally different applications of optimal transport.

---

9. Towards meta-pruning via optimal transport

URL: View paper

Brief Assessment

Meta-Pruning Transport[74] focuses on pruning neural networks by fusing neurons within a single model using optimal transport, not on merging heterogeneous models with different architectures. The paper addresses intra-model compression rather than cross-architecture model merging.

Contribution 3: Two-stage compression curriculum with layer-aware chunking

Description: The authors propose a training strategy that first learns a high-capacity latent representation using a deterministic autoencoder, then enables the KL term to structure the latent space. This curriculum, combined with layer-aware chunking of weight tensors, improves stability and out-of-distribution generalization when encoding LLM weights with heavy-tailed distributions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Learning Unified User Quantized Tokenizers for User Representation

URL: View paper

Brief Assessment

Unified User Tokenizers[59] focuses on user representation learning through a two-stage architecture for tokenization (embedding then discretization), not weight compression for LLM merging. The technical domains are fundamentally different.

---

2. 3D Skull Completion via Two-stage Conditional Diffusion-Based Signed Distance Fields

URL: View paper

Brief Assessment

Skull Completion Diffusion[56] focuses on 3D shape completion for cranial implants using VQ-VAE for spatial geometry encoding, not on weight compression for language models. The two-stage approach in the candidate refers to coarse-to-fine geometric reconstruction, fundamentally different from the original's curriculum for encoding LLM weight distributions.

---

3. Edge-Aware Reparameterizable Network With Hybrid Bottlenecks and Spatial Attention for Efficient Compressed Image Super-Resolution

URL: View paper

Brief Assessment

The candidate paper focuses on compressed image super-resolution using a two-stage training curriculum for image quality enhancement, not weight compression or variational autoencoders for LLM parameters. The training stages involve uncompressed bicubic data followed by JPEG/WebP/AVIF degradations, which is fundamentally different from the original paper's VAE-based weight encoding curriculum.

---

4. Map-Assisted Remote-Sensing Image Compression at Extremely Low Bitrates

URL: View paper

Brief Assessment

Map-Assisted Compression[57] focuses on remote-sensing image compression using diffusion models with a two-stage pipeline for image-to-latent mapping and conditional generation. This is fundamentally different from the original paper's two-stage curriculum for training VAEs on LLM weight tensors with layer-aware chunking to handle heavy-tailed distributions.

---

5. Surgical Robot Learning: From Demonstration and Simulation to World Models-A Review

URL: View paper

Brief Assessment

Surgical Robot Learning[60] mentions two-stage curriculum learning and CVAE training in robotics contexts, but does not address weight compression, latent space encoding of LLM parameters, or the specific heavy-tailed distribution challenges that motivate the original paper's approach.

---

6. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis

URL: View paper

Brief Assessment

RAVE[55] focuses on audio waveform synthesis using a two-stage training procedure (representation learning and adversarial fine-tuning) for VAEs, not on weight compression or LLM parameter encoding. The original paper's contribution addresses encoding neural network weights with heavy-tailed distributions using layer-aware chunking, which is fundamentally different from RAVE's audio signal processing domain.

---

7. Reducio! generating 1k video within 16 seconds using extremely compressed motion latents

URL: View paper

Brief Assessment

Reducio[54] focuses on video compression using a two-stage VAE training strategy (deterministic autoencoder followed by KL regularization) for motion latents, not LLM weight compression. The technical domain and application are fundamentally different from encoding LLM weights with heavy-tailed distributions.

---

8. Diffusion models for 3D generation: A survey

URL: View paper

Brief Assessment

Diffusion 3D Generation[51] focuses on 3D generation using diffusion models for point clouds and voxel grids, not on variational autoencoders for LLM weight compression. The two-stage curriculum mentioned applies to a completely different domain and task.

---

9. Rynnvla-001: Using human demonstrations to improve robot manipulation

URL: View paper

Brief Assessment

Rynnvla[53] focuses on robot manipulation using video generation pretraining and ActionVAE for action representation, not on variational autoencoders for weight compression or two-stage curriculum training for encoding LLM weights.

---

10. Machine Learning and Finite Element Simulation for Performance-Driven Generative Design in Aerodynamic Applications

URL: View paper

Brief Assessment

Performance-Driven Generative Design[52] focuses on aerodynamic design using machine learning and finite element methods, not on variational autoencoders for weight compression or two-stage curriculum training for neural network parameters.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] LS-Merge: Merging Language Models in Latent Space View paper
- [1] Mergenet: Knowledge migration across heterogeneous models, tasks, and modalities View paper
- [2] Knowledge fusion of large language models View paper
- [3] Amalgamating Multi-Task Models with Heterogeneous Architectures View paper

- [4] DeepMLF: Multimodal language model with learnable tokens for deep fusion in sentiment analysis View paper
- [5] Ensemble Learning for Heterogeneous Large Language Models with Deep Parallel Collaboration View paper
- [6] AdaMMS: Model Merging for Heterogeneous Multimodal Large Language Models with Unsupervised Coefficient Optimization View paper
- [7] Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities View paper
- [8] MergeME: Model Merging Techniques for Homogeneous and Heterogeneous MoEs View paper
- [9] Mixture-of-Agents Enhances Large Language Model Capabilities View paper
- [10] AudioPaLM: A Large Language Model That Can Speak and Listen View paper
- [11] Decision-Making Large Language Model for Wireless Communication: A Comprehensive Survey on Key Techniques View paper
- [12] LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers View paper
- [13] Spatiotemporal Pretrained Large Language Model for Forecasting With Missing Values View paper
- [14] Text2BIM: Generating Building Models Using a Large Language Model-Based Multiagent Framework View paper
- [15] Democratizing AI through model fusion: A comprehensive review and future directions View paper
- [16] Evolutionary optimization of model merging recipes View paper
- [17] Hierarchical Large Language Models in Cloud-Edge-End Architecture for Heterogeneous Robot Cluster Control View paper
- [18] Untangling the Influence of Typology, Data and Model Architecture on Ranking Transfer Languages for Cross-Lingual POS Tagging View paper
- [19] Effective and Explainable Molecular Property Prediction by Chain-of-Thought Enabled Large Language Models and Multi-Modal Molecular Information Fusion View paper
- [20] Sign Language Translation using Hybrid Temporal Convolutional Network with TTS Fusion View paper
- [21] LLM-WFIN: A Fine-Grained Large Language Model (LLM)-Oriented Website Fingerprinting Attack via Fusing Interrupt Trace and Network Traffic View paper
- [22] GeoLangBind: Unifying Earth Observation with Agglomerative Vision-Language Foundation Models View paper
- [23] Deep model fusion: A survey View paper
- [24] StatsMerging: Statistics-Guided Model Merging via Task-Specific Teacher Distillation View paper
- [25] Twin-Merging: Dynamic Integration of Modular Expertise in Model Merging View paper
- [26] (In) forming the new building envelope: A pedagogical study in generative design with precedents and multimodal large language models View paper
- [27] A survey on intelligent network operations and performance optimization based on large language models View paper
- [28] LaMoSC: Large Language Model-Driven Semantic Communication System for Visual Transmission View paper
- [29] Ethereum Fraud Detection via Joint Transaction Language Model and Graph Representation Learning View paper
- [30] SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation View paper
- [31] Autism Spectrum Disorder Diagnosis Through Natural Language Processing and ConvCapsule Fusion Network View paper
- [32] Fusechat: Knowledge fusion of chat models View paper
- [33] Multi-scale feature fusion and graph neural network integration for text classification with large language models View paper
- [34] Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models View paper
- [35] RigorLLM: Resilient Guardrails for Large Language Models against Undesired Content View paper
- [36] Extend Model Merging from Fine-Tuned to Pre-Trained Large Language Models via Weight Disentanglement View paper
- [37] Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts? View paper
- [38] Command A: An Enterprise-Ready Large Language Model View paper
- [39] MambaLLM: Integrating Macro-Index and Micro-Stock Data for Enhanced Stock Price Prediction View paper
- [40] SAM2CLIP2SAM: Vision Language Model for Segmentation of 3D CT Scans for Covid-19 Detection View paper
- [41] DAFNet: Dynamic Auxiliary Fusion for Sequential Model Editing in Large Language Models View paper
- [42] CoMoE: Collaborative Optimization of Expert Aggregation and Offloading for MoE-based LLMs at Edge View paper
- [43] Mixture of experts in large language models View paper
- [44] Artificial intelligence applications in education: Natural language processing in detecting misconceptions View paper
- [45] Merging and processing heterogeneous models View paper
- [46] Mitigating Catastrophic Forgetting in Language Transfer via Model Merging View paper
- [47] Dual-Branch Knowledge Enhancement Network with Vision-Language Model for Human-Object Interaction Detection View paper
- [48] GCN-LSTM: multi-label educational emotion prediction based on graph convolutional network and long and short term memory network fusion label correlation in â□¦ View paper
- [49] Model Assembly Learning with Heterogeneous Layer Weight Merging View paper
- [50] Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models View paper
- [51] Diffusion models for 3D generation: A survey View paper
- [52] Machine Learning and Finite Element Simulation for Performance-Driven Generative Design in Aerodynamic Applications View paper
- [53] Rynnvla-001: Using human demonstrations to improve robot manipulation View paper
- [54] Reducio! generating 1k video within 16 seconds using extremely compressed motion latents View paper
- [55] RAVE: A variational autoencoder for fast and high-quality neural audio synthesis View paper
- [56] 3D Skull Completion via Two-stage Conditional Diffusion-Based Signed Distance Fields View paper
- [57] Map-Assisted Remote-Sensing Image Compression at Extremely Low Bitrates View paper
- [58] Edge-Aware Reparameterizable Network With Hybrid Bottlenecks and Spatial Attention for Efficient Compressed Image Super-Resolution View paper
- [59] Learning Unified User Quantized Tokenizers for User Representation View paper
- [60] Surgical Robot Learning: From Demonstration and Simulation to World Models-A Review View paper
- [61] Emergent semantic entanglement in large language models: Non-sequential contextual weaving through stochastic syntagmatic bridges View paper
- [62] SeMe: Training-Free Language Model Merging via Semantic Alignment View paper
- [63] Latent syntax weaving in large language model representations: A novel mechanism for self-referential consistency in neural architectures View paper

- [64] Flows and Diffusions on the Neural Manifold View paper
- [65] EnergyMogen: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space View paper
- [66] Latent feature transformation for emergent task performance in large language models View paper
- [67] Unsupervised Neural Machine Translation with Weight Sharing View paper
- [68] Hierarchical Contextual Manifold Alignment for Structuring Latent Representations in Large Language Models View paper
- [69] EvoEdit: Lifelong Free-Text Knowledge Editing through Latent Perturbation Augmentation and Knowledge-driven Parameter Fusion View paper
- [70] AlignMerge - Alignment-Preserving Large Language Model Merging via Fisher-Guided Geometric Constraints View paper
- [71] Transformer fusion with optimal transport View paper
- [72] Model fusion via optimal transport View paper
- [73] Template based Graph Neural Network with Optimal Transport Distances View paper
- [74] Towards meta-pruning via optimal transport View paper
- [75] Merging embedded topics with optimal transport for online topic modeling on data streams View paper
- [76] Graph optimal transport for cross-domain alignment View paper
- [77] SuperGlue: Learning Feature Matching With Graph Neural Networks View paper
- [78] Unsupervised Learning for Optimal Transport plan prediction between unbalanced graphs View paper
- [79] A Survey on Optimal Transport for Machine Learning: Theory and Applications View paper
- [80] Fusion of Graph Neural Networks via Optimal Transport View paper