Harnessing Quantum Principles for Parameter-Efficient Continual Learning

Xiaobing Yu[1*] **Jin Yang**[1*] **Weiwei Ma**[1*] **Peijie Qiu**[1*]
Pan Xiao[1*] **Xiao Wu**[2] **Xiaofeng Liu**[3]

[1] Washington University in St. Louis
[2] Mohamed bin Zayed University of Artificial Intelligence
[3] Yale University

Abstract

Continual learning with parameter-efficient methods like LoRA prevents catastrophic forgetting but sacrifices cross-task knowledge transfer by freezing previous adapters. We recognize this mirrors quantum mechanics: how can multiple states coexist and interact? Qu-LoRA models task-specific LoRA adapters as quantum states in superposition, translating three quantum principles into concrete mechanisms: (1) superposition enables task coexistence through phase-controlled interference; (2) entanglement determines gradient sharing between related tasks while protecting unrelated ones; (3) measurement collapse eliminates task identity requirements, where inputs naturally select relevant knowledge through interference patterns. Unlike frozen approaches, Qu-LoRA achieves the impossible: previous tasks improve from subsequent learning while reducing forgetting by 75%. Experiments demonstrate superior performance across benchmarks, establishing quantum mechanics as a powerful CL framework.

1 Introduction

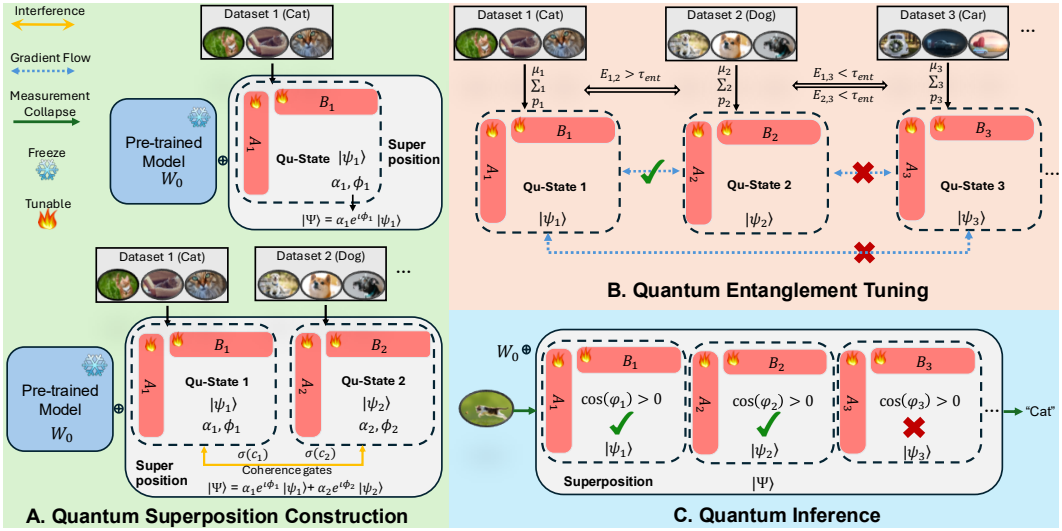
Continual learning (CL) remains one of the fundamental challenges in deep learning, expecting models to sequentially acquire new knowledge while preserving previously learned capabilities. It becomes particularly critical when adapting large-scale pre-trained models, where we must preserve both the rich representations learned during pretraining and task-specific knowledge acquired sequentially. Despite significant progress, existing approaches face a fundamental trade-off; they either freeze parameters to prevent forgetting, but sacrifice plasticity [20, 15, 14, 13], or allow parameter updates that risk catastrophic interference [6, 19]. Recent advances in parameter-efficient fine-tuning (PEFT), particularly Low-Rank Adaptation (LoRA) [4], have shown promise for CL on pre-trained models by isolating task-specific knowledge in small adapter modules while keeping pre-trained weights frozen. However, current LoRA-based CL methods universally adopt the freezing paradigm, once a task is learned, its adapters become immutable. Although freezing effectively prevents catastrophic forgetting, this comes with significant limitations: early tasks remain stuck with potentially suboptimal solutions, related tasks cannot mutually reinforce each other's learning, and valuable insights from later tasks can never improve earlier ones [8, 18]. For instance, learning about "trucks" in Task 5 could refine Task 1's "cars" classifier, but frozen adapters make this impossible.

The core challenge of CL is that frozen adapters prevent catastrophic forgetting but cannot improve from future tasks, while unfrozen adapters enable improvement but suffer forgetting. We need frameworks where tasks can selectively share knowledge without requiring explicit task identity during inference. We present Quantum-inspired Low-Rank Adaptation (Qu-LoRA), the first CL

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI That Keeps Up: Workshop on Continual and Compatible Foundation Model Updates.

Figure 1: Our Qu-LoRA framework leverages three quantum-inspired mechanisms. **(A)**. Quantum superposition allows task adapters to coexist with complex amplitudes and phases. **(B)**. Quantum entanglement determines task relationships and controls gradient flow between adapters. **(C)**. During inference, quantum measurement collapse naturally selects relevant knowledge through interference patterns without requiring task identity.



framework enabling non-frozen LoRA adapters. Drawing from quantum mechanics, where states exist in superposition, become entangled, and collapse upon measurement. We implement three mechanisms: (1) Quantum superposition: representing each task adapter as a quantum state with learned amplitudes and phases, enabling coexistence through interference patterns; (2) Quantum entanglement: determining task relationships via phase assignment algorithms based on feature distribution similarity, allowing related tasks to share beneficial gradient updates while isolating unrelated ones; (3) Schrödinger's cat-inspired [5] inference: the model maintains superposition of all task states until prediction triggers natural collapse to the most relevant knowledge, eliminating the need for explicit task identity.

Our approach represents a paradigm shift from traditional CL methods. While regularization-based, replay-based [11], and architecture-based [17] approaches fundamentally assume that protecting old knowledge requires restricting parameter updates, and prompt-based methods [13] isolate tasks through separate tokens, Qu-LoRA demonstrates that all parameters can remain adaptable while maintaining stability. The quantum framework provides not just inspiration but a complete practical toolkit. Our main contributions include:

• We introduce the first LoRA-based CL framework that never freezes adapters, solving a fundamental limitation where early tasks cannot benefit from later learning. Our quantum-inspired approach enables backward knowledge transfer while preventing catastrophic forgetting.

• We develop a controlled gradient flow mechanism guided by quantum entanglement that allows only beneficial updates between related tasks. This enables mutual improvement while protecting unrelated tasks from interference, reducing forgetting by 75% compared to frozen baselines.

• We eliminate task identity requirements through quantum superposition inference, where all task knowledge remains simultaneously accessible with phase-based interference: related tasks reinforce each other through constructive interference while unrelated tasks actively cancel through destructive interference until measurement collapse naturally selects the relevant adaptation.

• Our proposed method achieves superior performance across multiple benchmarks: 91.17% on Split CIFAR-100 [7], 79.87% on Split ImageNet-R [3] with up to 75% reduction in forgetting versus frozen baselines and maintains high performance.

2 Methodology

Quantum mechanics offers a different perspective by providing theoretical foundations to control interaction precisely rather than preventing it. We translate three quantum principles into neural network operations. Task parameters become quantum states with complex amplitudes, enabling controlled interference through phase relationships. Statistical task similarity becomes quantum entanglement, determining gradient flow between tasks. Input features act as measurement operators that collapse the superposition, where learned phases cause related tasks to reinforce while unrelated tasks cancel out.

It enables Qu-LoRA to achieve what existing methods cannot: backward knowledge transfer without catastrophic forgetting through non-frozen adapters.

Quantum Superposition: Task Coexistence Each task's LoRA adaptation exists as a quantum state in superposition with others:

$$|\Psi\rangle = \sum_{i=0}^{n-1} \alpha_i e^{\iota\phi_i} |\psi_i\rangle \tag{1}$$

where α_i controls task contribution strength and ϕ_i determines interference patterns. Unlike classical averaging where all tasks contribute positively, the phase term $\cos(\phi_i)$ enables both constructive and destructive interference, allowing multiple non-frozen adapters to coexist without conflict.

Quantum Entanglement: Selective Knowledge Transfer We measure task relationships through multi-order statistical similarity (mean alignment, covariance correlation, and distributional distance) to compute entanglement scores E_{ij}. Highly entangled tasks receive similar phases for mutual reinforcement, while independent tasks receive opposite phases for isolation. Related tasks share beneficial updates through positive interference coefficients, while unrelated tasks remain protected through negative interference. This controlled gradient mechanism enables backward knowledge transfer, where early tasks improve from later learning.

Schrödinger's Cat Inference: Task-Free Prediction During inference, the model maintains all tasks in superposition until the input "measures" the system:

$$y = \sum_{i=0}^{n-1} \alpha_i \cos(\phi_i)\sigma(c_i) f_i(x) \tag{2}$$

The interference term $\cos(\phi_i)$ naturally selects relevant knowledge: related tasks reinforce (positive cosine), unrelated tasks cancel (negative cosine). This eliminates the need for explicit task identity, where the input itself causes the superposition to collapse to the appropriate knowledge through quantum interference (More detail can be found in Appendix).

Qu-LoRA is the **first LoRA-based CL framework with permanently non-frozen adapters**, achieving backward knowledge transfer without catastrophic forgetting. It attains 91.17% on Split CIFAR-100, 79.87% on Split ImageNet-R, with a 75% reduction in forgetting compared to frozen baselines, while requiring no task identification during deployment.

3 Experiments

We train and evaluate Qu-LoRA on multiple challenging CL benchmarks following the class-incremental setting: Split CIFAR-100 (100 classes) and Split ImageNet-R (200 classes of artistic renditions) to test multi-domain CL.

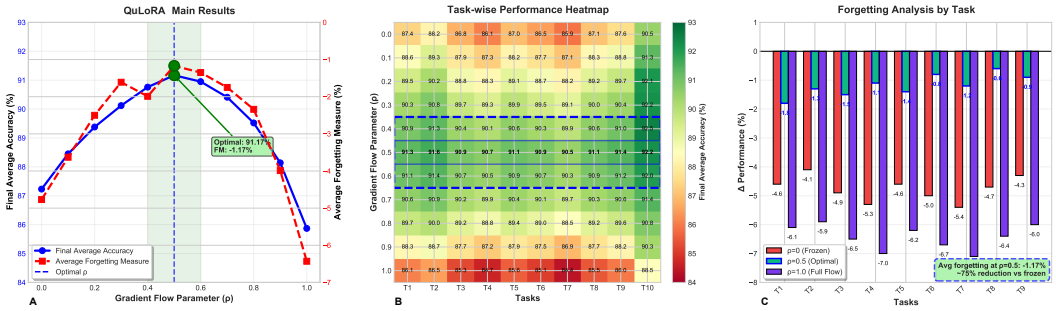
As detailed in Tab.1, we systematically compared Qu-LoRA with representative methods from four CL paradigms based on regularization, prompt, adapter, and LoRA, respectively. All methods use the same ViT-B/16 backbone with ImageNet-21K pre-training [1].

Following standard evaluation protocols in CL, we adopt three key metrics for comprehensive assessment. Final Average Accuracy (FA) measures the model's overall performance across all tasks upon completion of training. Cumulative Average Accuracy (CA) tracks the evolution of performance by computing the running average throughout the learning sequence. Average Forgetting Measure

Table 1: Overall performance comparison on Split CIFAR-100 and Split ImageNet-R in 10 incremental tasks. We present Final Average Accuracy (FA%), Cumulative Average Accuracy (CA%), and Average Forgetting Measure (FM%) of all methods under different CL methods on the same pre-trained model (**ViT-B/16-IN21K**).

Method Category	Method	Split CIFAR-100			Split ImageNet-R		
		FA (↑)	**CA**(↑)	**FM**(↓)	**FA** (↑)	**CA**(↑)	**FM**(↓)
Fine-tune		34.17 ± 7.55	37.51 ± 8.20	58.37 ± 9.63	31.71 ± 6.77	36.69 ± 8.37	60.51 ± 13.51
Regularization	EWC [6]	86.01 ± 1.86	89.42 ± 1.71	4.57 ± 0.56	70.31 ± 1.17	75.80 ± 0.52	5.72 ± 1.40
Prompt-based	L2P [15]	80.71 ± 1.47	86.61 ± 0.91	6.29 ± 1.17	65.61 ± 0.75	69.28 ± 0.81	8.27 ± 1.72
	HiDe-Prompt [13]	89.55 ± 0.28	91.61 ± 0.21	2.01 ± 0.28	72.31 ± 0.21	74.17 ± 0.15	5.36 ± 0.27
Adapter-based	SEMA [12]	85.17 ± 0.35	88.25 ± 0.44	2.46 ± 0.51	78.56 ± 0.33	79.95 ± 0.48	6.77 ± 0.58
	EASE [20]	88.32 ± 0.51	92.16 ± 0.21	1.94 ± 0.27	75.26 ± 1.01	78.57 ± 0.45	4.98 ± 0.81
LoRA-based	InfLoRA [8]	82.17 ± 0.49	86.88 ± 0.72	3.56 ± 0.36	72.44 ± 0.74	73.60 ± 0.83	7.15 ± 0.72
	SD-LoRA [16]	87.16 ± 1.92	89.94 ± 0.69	1.69 ± 1.98	77.48 ± 0.68	79.73 ± 1.07	5.36 ± 0.67
	CL-LoRA [2]	87.31 ± 0.18	90.22 ± 0.16	1.96 ± 0.13	75.84 ± 0.29	77.84 ± 0.96	5.96 ± 0.78
	Ours	**91.17** ± 0.09	**93.00** ± 0.31	**1.17** ± 0.24	**79.87** ± 0.27	**81.95** ± 0.46	**3.10** ± 0.39

Figure 2: Qu-LoRA performance analysis on CIFAR-100. **(A)** Effect of gradient flow parameter ρ on FA and FM; **(B)** Task-wise accuracy heatmap where each cell represents the final accuracy of all tasks with corresponding ρ, with diagonal values showing immediate task performance and off-diagonal values revealing knowledge retention; **(C)** Per-task forgetting comparison between different ρ. The controlled gradient flow ($\rho = 0.5$) achieves lower forgetting than frozen adapters ($\rho = 0$).



(FM) quantifies the extent of knowledge degradation on previous tasks. Among them, we prioritize FA and CA as they provide the most holistic view of CL performance, capturing both final capabilities and learning dynamics throughout the training process.

Performance on Benchmark Datasets. In Tab.1, we demonstrate Qu-LoRA's effectiveness by achieving 91.17% FA with 1.17% average forgetting on Split CIFAR-100, outperforming HiDe-Prompt. On ImageNet-R, Qu-LoRA reaches 79.87% FA with 3.10% FM. We observe significantly improved knowledge retention through controlled gradient flow, with Qu-LoRA showing only 1.17% average forgetting compared to 4.65% for frozen LoRA baselines.

Gradient Flow Analysis: Fig.2.A reveals optimal gradient flow at $\rho = 0.5$, achieving 91.17% accuracy with only 1.17% forgetting on CIFAR-100, a 75% reduction from frozen adapters (4.65%). This validates our core hypothesis: controlled gradient flow enables positive backward transfer. In Fig.2.B, the task-wise heatmap shows Qu-LoRA maintains >85% performance across all task pairs, unlike traditional methods with degrading off-diagonal values. Remarkably, in Fig.2.C, several tasks exhibit minimal forgetting, with significantly reduced performance degradation compared to frozen architectures. This demonstrates that controlled gradient flow dramatically reduces forgetting while maintaining strong performance across all tasks.

4 Conclusion

We presented Qu-LoRA, the first CL framework that enables non-frozen LoRA adapters through quantum-inspired mechanisms, achieving backward knowledge transfer without catastrophic forgetting. Our quantum framework fundamentally shifts the CL paradigm from isolation to controlled interaction, opening new possibilities for adaptive learning systems that continuously improve across all tasks.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] J. He, Z. Duan, and F. Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30534–30544, 2025.

[3] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[5] S. C. Kak. Quantum neural computing. volume 94 of *Advances in Imaging and Electron Physics*, pages 259–313. Elsevier, 1995.

[6] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[7] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[8] Y.-S. Liang and W.-J. Li. Inflora: Interference-free low-rank adaptation for continual learning. In *CVPR*, pages 23638–23647, 2024.

[9] M.-L. Luo, Z.-H. Zhou, T. Wei, and M.-L. Zhang. Lada: Scalable label-specific clip adapter for continual learning. *arXiv preprint arXiv:2505.23271*, 2025.

[10] D. Miki, A. Matsumura, and K. Yamamoto. Non-gaussian entanglement in gravitating masses: The role of cumulants. *Physical Review D*, 105(2):026011, 2022.

[11] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017.

[12] H. Wang, H. Lu, L. Yao, and D. Gong. Self-expansion of pre-trained models with mixture of adapters for continual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10087–10098, 2025.

[13] L. Wang, J. Xie, X. Zhang, M. Huang, H. Su, and J. Zhu. Hierarchical decomposition of prompt-based continual learning: Rethinking obscured sub-optimality. *Neurips*, 36, 2024.

[14] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pages 631–648. Springer, 2022.

[15] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022.

[16] Y. Wu, H. Piao, L.-K. Huang, R. Wang, W. Li, H. Pfister, D. Meng, K. Ma, and Y. Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *ICLR*, 2025.

[17] S. Yan, J. Xie, and X. He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, pages 3014–3023, 2021.

[18] X. Yu, J. Yang, X. Wu, P. Qiu, and X. Liu. Fm-lora: Factorized low-rank meta-prompting for continual learning. In *CVPR Workshop*, pages 6409–6418, 2025.

[19] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

[20] D.-W. Zhou, H.-L. Sun, H.-J. Ye, and D.-C. Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *CVPR*, pages 23554–23564, 2024.

A More details for Methology

Our Qu-LoRA enables non-frozen adapters through quantum mechanical principles. To achieve this, we must solve three major technical challenges. First, how can multiple task parameters coexist without destructive interference? Second, how do we determine which tasks should share gradients and by how much? Third, how can the model leverage all learned knowledge during inference when task identity are unknown and asking "which task is this?" is analogous to ask whether Schrödinger's cat is alive or dead before opening the box?

Traditional solutions use isolated modules, frozen parameters, or explicit routing. Quantum mechanics offers a different perspective by providing theoretical foundations to control interaction precisely rather than preventing it. We translate three quantum principles into neural network operations. Task parameters become quantum states with complex amplitudes, enabling controlled interference through phase relationships. Statistical task similarity becomes quantum entanglement, determining gradient flow between tasks. Input features act as measurement operators that collapse the superposition, where learned phases cause related tasks to reinforce while unrelated tasks cancel out.

It enables Qu-LoRA to achieve what existing methods cannot: backward knowledge transfer without catastrophic forgetting through non-frozen adapters.

A.1 Quantum Superposition: Enabling Task Coexistence

To enable non-frozen adapters without destructive interference, we model each task's knowledge as a quantum state that can exist in superposition with others. For a pre-trained model with frozen weights $W_0 \in \mathbb{R}^{d\times k}$ where d is the output dimension and k is the input dimension, each task i learns low-rank adaptation matrices $A_i \in \mathbb{R}^{d\times r}$ and $B_i \in \mathbb{R}^{r\times k}$ with rank $r \ll \min(d,k)$. The weight update for task i is $\Delta W_i = A_i B_i$ for $W_i = W_0 + \Delta W_i$.

We encode each task's adaptation as a normalized quantum state $|\psi_i\rangle = \text{vec}(A_i B_i)/||\text{vec}(A_i B_i)||$, where $\text{vec}(\cdot) : \mathbb{R}^{d\times k} \to \mathbb{R}^{dk}$ denotes vectorization that stacks columns into a single vector. The normalization ensures unit norm for quantum state representation (formal Hilbert space construction in **Suppl. Mat.**)

The complete system exists in a quantum superposition:

$$|\Psi\rangle = \sum_{i=0}^{n-1} \alpha_i e^{\iota\phi_i} |\psi_i\rangle. \tag{3}$$

Here, n denotes the total number of tasks learned so far, and $\iota = \sqrt{-1}$ is the imaginary unit. The complex amplitude $\alpha_i e^{\iota\phi_i}$ consists of two components: $\alpha_i \in \mathbb{R}^+$ is a learnable amplitude controlling task i's contribution strength (larger α_i means greater influence during inference), and $\phi_i \in [0, 2\pi)$ is the quantum phase determining interference patterns with other tasks. The imaginary unit ι is essential for quantum superposition as it enables complex-valued states that can exhibit interference, when we extract the real part via $\cos(\phi_i)$, tasks can constructively reinforce (positive values) or destructively cancel (negative values) based on their phase relationships. The amplitudes satisfy Born rule normalization $\sum_{i=0}^{n-1} |\alpha_i|^2 = 1$, ensuring probabilistic interpretation where $|\alpha_i|^2$ represents the probability of observing task i upon measurement (Detailed proof in **Suppl. Mat.**).

Theorem 1 (Quantum Superposition in Parameter Space) *During training and forward propagation, we do NOT collapse this superposition. The system evolves according to quantum mechanics, maintaining all tasks in coherent superposition. The observable quantities are computed through quantum expectation values:*

$$\langle \Delta W \rangle = \langle \Psi | \hat{W} | \Psi \rangle. \tag{4}$$

Only during inference does measurement collapse occur:

$$y = \sum_{i=0}^{n-1} \alpha_i e^{\iota\phi_i} \sigma(c_i) \langle x|\psi_i\rangle f_i(x), \tag{5}$$

where the input x acts as measurement operator, causing the superposition to collapse based on the quantum mechanical overlap $\langle x|\psi_i\rangle$.

The key innovation is the phase-dependent interference term $\cos(\phi_i)$. Unlike classical parameter averaging $\Delta W_{\text{classical}} = \sum_{i=0}^{n-1} |\alpha_i|^2 (A_i B_i)$ where all tasks contribute positively with probability weights $|\alpha_i|^2$, our quantum formulation uses amplitudes α_i and allows both constructive and destructive interference. When two tasks have similar phases such that $\phi_i \approx \phi_j$, then $\cos(\phi_i) \approx \cos(\phi_j)$ and both contribute positively, reinforcing each other. When phases differ by π such that $\phi_i \approx \phi_j + \pi$, then $\cos(\phi_i) \approx -\cos(\phi_j)$ and they cancel out. This is essential for multiple non-frozen adapters to coexist.

The quantum parameters evolve during training as follows. Amplitudes α_i are optimized to maximize task performance on validation data through gradient ascent: $\alpha_i \leftarrow \alpha_i + \eta_\alpha \nabla_{\alpha_i} \mathcal{L}_{\text{val}}$ where $\mathcal{L}_{\text{val}}$ is the validation loss and η_α is the learning rate. Phases ϕ_i are determined by the entanglement structure as described in later Eq.(18). The coherence gates $\sigma(c_i)$ are learned through standard backpropagation to balance quantum coherence with training stability.

To maintain tasks in quantum superposition and enable proper interference, we introduce the coherence preservation regularization:

$$\mathcal{L}_{\text{coherence}} = \lambda_c \sum_{i=0}^{n-1} (1 - \sigma(c_i))^2, \tag{6}$$

where λ_c is the regularization strength. This penalizes low coherence values, forces $\sigma(c_i) \approx 1$, so previous tasks remain in superposition and can interfere quantum-mechanically during both training and inference. Without this regularization, tasks would collapse to classical states, losing the interference capabilities essential for selective knowledge sharing.

A.2 Quantum Entanglement: Enabling Backward Knowledge Transfer with Non-Frozen Adapters

All existing LoRA-based CL methods freeze adapters because unrestricted gradient flow causes catastrophic forgetting, which previous tasks' knowledge gets overwritten when learning new tasks [6, 9]. To enable backward knowledge transfer where early tasks can benefit from later learning, we need a principled way to control which tasks share gradients and how much.

Quantum entanglement provides the theoretical framework for solving this dilemma. In which entangled particles can share information instantaneously while maintaining their individual quantum states. This principle maps perfectly to our challenge: we need tasks to share beneficial knowledge while protecting individual learned representations.

Connection to Quantum Superposition. In previous section, we established that tasks exist in superposition with complex amplitudes $\alpha_i e^{\iota \phi_i}$. However, superposition alone doesn't determine which tasks should interfere constructively or destructively, this is where entanglement becomes crucial. Entanglement determines the phase relationships: highly entangled tasks receive similar phases for constructive interference, while independent tasks receive opposite phases for destructive interference. This creates a complete quantum system where superposition enables coexistence and entanglement controls interactions.

Multi-Order Task Entanglement. While quantum entanglement is traditionally measured through correlated measurement outcomes, we adapt this concept to neural networks by measuring statistical correlations in feature space. This parallels quantum optics, where first-order coherence functions measure field correlations (analogous to our mean feature alignment), while higher-order coherence functions reveal intensity correlations and quantum statistics (analogous to our covariance and distributional measures) [10]. Just as distinguishing between thermal, coherent, and non-classical light requires multiple coherence orders, identifying truly entangled tasks requires multi-order statistical analysis:

First-Order Entanglement (Mean Feature Alignment) captures whether tasks recognize similar high-level patterns:

$$F(\mu_i, \mu_j) = \frac{1}{2} \left[1 + \frac{\langle \mu_i, \mu_j \rangle}{||\mu_i|| \cdot ||\mu_j||} \right], \tag{7}$$

where $\mu_i \in \mathbb{R}^d$ is the average feature vector at the penultimate layer for task i. When $F(\mu_i, \mu_j) \approx 1$, tasks process similar semantic information (e.g.,"cats" and "dogs" both activate animal-related features).

Second-Order Entanglement (Correlated Feature Variation) measures whether tasks vary along similar feature dimensions:

$$S(\Sigma_i, \Sigma_j) = \frac{1}{2}\left[1 + \frac{\langle \text{vec}(\Sigma_i), \text{vec}(\Sigma_j)\rangle}{||\text{vec}(\Sigma_i)|| \cdot ||\text{vec}(\Sigma_j)||}\right], \tag{8}$$

where $\Sigma_i \in \mathbb{R}^{d\times d}$ is the covariance matrix. This captures whether tasks not only have similar means but also similar variation patterns, which is crucial for identifying truly related tasks.

Higher-Order Entanglement (Overall Distributional Similarity) captures complete statistical relationships using symmetrized KL divergence for distributional differences:

$$D_{KL}^{\text{sim}}(p_i||p_j) = \exp\left(-\frac{D_{KL}(p_i||p_j) + D_{KL}(p_j||p_i)}{2\lambda}\right). \tag{9}$$

The multi-order entanglement score aggregates these measures:

$$E_{ij} = \frac{1}{3}[F(\mu_i, \mu_j) + S(\Sigma_i, \Sigma_j) + D_{KL}^{\text{sim}}(p_i||p_j)]. \tag{10}$$

Entanglement Controls Phase Assignment. The entanglement score directly determines phase relationships in our quantum superposition. High entanglement indicates tasks share similar feature statistics across all orders, they recognize similar patterns, vary along similar dimensions, and have similar distributions. Low entanglement means tasks operate on entirely different feature spaces with minimal statistical overlap. During phase assignment, we ensure:

$$\text{If } E_{ij} \approx 1 \text{ (highly entangled)} \Rightarrow \phi_i \approx \phi_j. \tag{11}$$

$$\text{If } E_{ij} \approx 0 \text{ (independent)} \Rightarrow \phi_i \approx \phi_j + \pi. \tag{12}$$

This creates a coherent quantum system where entanglement determines how tasks interact in superposition.

Quantum-Controlled Gradient Flow. The entanglement score not only determines phases but also controls gradient flow, implementing selective information sharing. During training of task n, the forward pass incorporates previous tasks based on entanglement:

$$y_n(x) = f_n(x) + \sum_{i=0}^{n-1} \alpha_{in}(E_{in}) f_i(x). \tag{13}$$

The interference coefficient $\alpha_{in}(E_{in}) = \beta \cdot \tanh(k(E_{in} - \tau_{\text{ent}})) \cdot E_{in}$ creates positive interference ($\alpha_{in} > 0$) for entangled tasks, enabling knowledge sharing, and negative interference ($\alpha_{in} < 0$) for independent tasks, preventing harmful transfer.

During backpropagation, this same entanglement controls gradient flow:

$$\frac{\partial \mathcal{L}}{\partial A_i} = \frac{\partial \mathcal{L}}{\partial y_n} \cdot \alpha_{in}(E_{in}) \cdot \rho \cdot \frac{\partial f_i}{\partial A_i}. \tag{14}$$

The complete quantum picture: tasks exist in superposition, entanglement determines their relationships and phase assignments, and measurement collapse selects relevant knowledge during inference). This creates a unified quantum framework where each component plays a specific role.

Quantum-Inspired Protection Against Forgetting. To ensure stability while allowing gradient flow to non-frozen adapters, we introduce regularization terms inspired by quantum mechanics:

• *Quantum Zeno Effect* protects stable parameters from drift:

$$\mathcal{L}_{\text{Zeno}} = \lambda_z \sum_{i=0}^{n-1} \sigma(c_i)||\theta_i - \theta_i^{(n-1)}||_2^2, \tag{15}$$

where protection strength is proportional to coherence $\sigma(c_i)$.

• *Entanglement Maintenance* preserves discovered task relationships:

$$\mathcal{L}_{\text{entangle}} = \lambda_e \sum_{i=0}^{n-1} \sum_{\substack{j=0 \\ j \neq i}}^{n-1} E_{ij}^2 \frac{||A_i B_i - A_j B_j||_F^2}{||A_i B_i||_F ||A_j B_j||_F}. \tag{16}$$

• *Uncertainty-Inspired Gradient Bounding* limits update magnitudes:

$$\mathcal{L}_{\text{measurement}} = \lambda_m \sum_{i=0}^{n-1} \text{ReLU}\left(\frac{\gamma}{||\theta_i|| + \epsilon} - ||\nabla_{\theta_i} \mathcal{L}_{\text{task}}^{(n)}||\right). \tag{17}$$

Through quantum entanglement, we achieve what existing methods cannot: selective knowledge sharing between related tasks while protecting unrelated ones, enabling true backward transfer without catastrophic forgetting.

A.3 Quantum Measurement Collapse: Task-Free Inference Through Interference

While existing methods achieve task-free inference through discrete selection mechanisms (prompt matching, learned routers, or weighted combinations), they cannot actively suppress irrelevant knowledge as all components contribute positively. Our quantum measurement collapse enables a fundamentally different mechanism: true interference-based selection where related tasks reinforce each other while unrelated tasks actively cancel out, eliminating the need for any task identification.

This mirrors Schrödinger's cat principle in quantum mechanics where the system exists in superposition of all possible states until measurement causes collapse to a definite outcome. Similarly, our model maintains all task knowledge in quantum superposition until the input "measures" the system, causing natural collapse through interference patterns rather than explicit selection.

Phase Assignment for Interference Control. The power of quantum interference lies in phase relationships. Before training each task n, we assign quantum phases that determine how tasks will interfere during inference:

$$\phi_i^{(n)} = \arg\min_{\phi} \sum_{j=0}^{n-1} \left(\cos(\phi - \phi_j^{(n)}) - \cos(\pi(1 - E_{ij}))\right)^2. \tag{18}$$

This optimization ensures highly entangled tasks (large E_{ij}) receive similar phases for constructive interference. When $E_{ij} \approx 1$, the target becomes $\cos(\phi - \phi_j) \approx 1$, achieved when $\phi \approx \phi_j$. Conversely, independent tasks (small E_{ij}) receive opposite phases for destructive interference. When $E_{ij} \approx 0$, the target becomes $\cos(\phi - \phi_j) \approx -1$, achieved when $\phi \approx \phi_j + \pi$. The phases remain fixed during training task t, ensuring stable quantum dynamics.

Quantum Superposition Inference. During inference, we eliminate task identity requirements through quantum superposition:

Theorem 2 (Schrödinger's Cat Principle) *During inference without task identity, the model exists in quantum superposition of all n learned task states, analogous to Schrödinger's cat being simultaneously alive and dead until observed. For any input x, all task adapters remain active in coherent superposition until the prediction "measurement" causes the system to naturally collapse to relevant task knowledge:*

$$y = \textit{Real}\left[\sum_{i=0}^{n-1} \alpha_i e^{\iota \phi_i^{(n)}} \sigma(c_i) \langle x | \psi_i \rangle f_i(x)\right] \tag{19}$$

$$= \sum_{i=0}^{n-1} \alpha_i \cos(\phi_i^{(n)}) \sigma(c_i) f_i(x), \tag{20}$$

where the measurement process extracts the real part of the quantum amplitude since $e^{\iota\phi} = \cos(\phi) + \iota \sin(\phi)$ and we take only the real component.

The key distinction from existing methods lies in the interference term $\cos(\phi_i^{(n)})$. Consider an input belonging to a “cats" task. Related tasks such as “dogs" have similar phases, so $\cos(\phi_{\text{dogs}}) > 0$ contributes positively. Unrelated tasks such as “cars" have opposite phases, so $\cos(\phi_{\text{cars}}) < 0$ contributes negatively. These negative contributions actively cancel irrelevant knowledge.

This is fundamentally different from methods like SD-LoRA that use $y = \sum \alpha_i f_i(x)$ where all $\alpha_i > 0$, or prompt-based methods that select discrete components. Our approach implements true quantum measurement collapse where the input characteristics naturally cause the superposition to collapse toward relevant knowledge through constructive and destructive interference. This enables active suppression of irrelevant tasks (impossible with positive-only weights), automatic blending for ambiguous inputs, and seamless scalability as new tasks simply add terms to the superposition. Our experiments validate this: Qu-LoRA achieves only 3.19% gap from oracle performance while requiring no task identification, making it genuinely practical for real-world deployment.