1601. ONLINE EPSILON NET & PIERCING SET FOR GEOMETRIC CONCEPTS

链接：https://iclr.cc/virtual/2025/poster/28409 abstract： VC-dimension (Vapnik & Chervonenkis (1971)) and ε-nets (Haussler & Welzl (1987)) are key concepts in Statistical Learning Theory. Intuitively, VC-dimension is a measure of the size of a class of sets. The famous ε-net theorem, a fundamental result in Discrete Geometry, asserts that if the VC-dimension of a set system is bounded, then a small sample exists that intersects all sufficiently large sets. In online learning scenarios where data arrives sequentially, the VC-dimension helps to bound the complexity of the set system, and ε-nets ensure the selection of a small representative set. This sampling framework is crucial in various domains, including spatial data analysis, motion planning in dynamic environments, optimization of sensor networks, and feature extraction in computer vision, among others. Motivated by these applications, we study the online ε-net problem for geometric concepts with bounded VC-dimension. While the offline version of this problem has been extensively studied, surprisingly, there are no known theoretical results for the online version to date. We present the first deterministic online algorithm with an optimal competitive ratio for intervals in $\mathbb{R}$. Next, we give a randomized online algorithm with a near-optimal competitive ratio for axis-aligned boxes in $\mathbb{R}^d$, for $d\le 3$. Furthermore, we introduce a novel technique to analyze similar-sized objects of constant description complexity in $\mathbb{R}^d$, which may be of independent interest. Next, we focus on the continuous version of this problem (called online piercing set), where ranges of the set system are geometric concepts in $\mathbb{R}^d$ arriving in an online manner, but the universe is the entire ambient space, and the objective is to choose a small sample that intersects all the ranges. Although online piercing set is a very well-studied problem in the literature, to our surprise, very few works have addressed generic geometric concepts without any assumption about the sizes. We advance this field by proposing asymptotically optimal competitive deterministic algorithms for boxes and ellipsoids in $\mathbb{R}^d$, for any $d\in\mathbb{N}$.

1602. Vision CNNs trained to estimate spatial latents learned similar ventral-stream-aligned representations

链接：https://iclr.cc/virtual/2025/poster/28907 abstract： Studies of the functional role of the primate ventral visual stream have traditionally focused on object categorization, often ignoring -- despite much prior evidence -- its role in estimating "spatial" latents such as object position and pose. Most leading ventral stream models are derived by optimizing networks for object categorization, which seems to imply that the ventral stream is also derived under such an objective. Here, we explore an alternative hypothesis: Might the ventral stream be optimized for estimating spatial latents? And a closely related question: How different -- if at all -- are representations learned from spatial latent estimation compared to categorization? To ask these questions, we leveraged synthetic image datasets generated by a 3D graphic engine and trained convolutional neural networks (CNNs) to estimate different combinations of spatial and category latents. We found that models trained to estimate just a few spatial latents achieve neural alignment scores comparable to those trained on hundreds of categories, and the spatial latent performance of models strongly correlates with their neural alignment. Spatial latent and category-trained models have very similar -- but not identical -- internal representations, especially in their early and middle layers. We provide evidence that this convergence is partly driven by non-target latent variability in the training data, which facilitates the implicit learning of representations of those non-target latents. Taken together, these results suggest that many training objectives, such as spatial latents, can lead to similar models aligned neurally with the ventral stream. Thus, one should not assume that the ventral stream is optimized for object categorization only. As a field, we need to continue to sharpen our measures of comparing models to brains to better understand the functional roles of the ventral stream.

1603. Online Reward-Weighted Fine-Tuning of Flow Matching with Wasserstein Regularization

链接：https://iclr.cc/virtual/2025/poster/31149 abstract： Recent advancements in reinforcement learning (RL) have achieved great success in fine-tuning diffusion-based generative models. However, fine-tuning continuous flow-based generative models to align with arbitrary user-defined reward functions remains challenging, particularly due to issues such as policy collapse from overoptimization and the prohibitively high computational cost of likelihoods in continuous-time flows. In this paper, we propose an easy-to-use and theoretically sound RL fine-tuning method, which we term Online Reward-Weighted Conditional Flow Matching with Wasserstein-2 Regularization (ORW-CFM-W2). Our method integrates RL into the flow matching framework to fine-tune generative models with arbitrary reward functions, without relying on gradients of rewards or filtered datasets. By introducing an online reward-weighting mechanism, our approach guides the model to prioritize high-reward regions in the data manifold. To prevent policy collapse and maintain diversity, we incorporate Wasserstein-2 (W2) distance regularization into our method and derive a tractable upper bound for it in flow matching, effectively balancing exploration and exploitation of policy optimization. We provide theoretical analyses to demonstrate the convergence properties and induced data distributions of our method, establishing connections with traditional RL algorithms featuring Kullback-Leibler (KL) regularization and offering a more comprehensive understanding of the underlying mechanisms and learning behavior of our approach. Extensive experiments on tasks including target image generation, image compression, and text-image alignment demonstrate the effectiveness of our method, where our method achieves optimal policy convergence while allowing controllable trade-offs between reward maximization and diversity preservation.

1604. Unsupervised Meta-Learning via In-Context Learning

链接：https://iclr.cc/virtual/2025/poster/30089 abstract： Unsupervised meta-learning aims to learn feature representations from

unsupervised datasets that can transfer to downstream tasks with limited labeled data.In this paper, we propose a novel approach to unsupervised meta-learning that leverages the generalization abilities of in-context learning observed in transformer architectures. Our method reframes meta-learning as a sequence modeling problem, enabling the transformer encoder to learn task context from support images and utilize it to predict query images. At the core of our approach lies the creation of diverse tasks generated using a combination of data augmentations and a mixing strategy that challenges the model during training while fostering generalization to unseen tasks at test time. Experimental results on benchmark datasets showcase the superiority of our approach over existing unsupervised meta-learning baselines, establishing it as the new state-of-the-art. Remarkably, our method achieves competitive results with supervised and self-supervised approaches, underscoring its efficacy in leveraging generalization over memorization.

1605. MaxInfoRL: Boosting exploration in reinforcement learning through information gain maximization

链接：https://iclr.cc/virtual/2025/poster/29662 abstract：

1606. AdaGrad under Anisotropic Smoothness

链接：https://iclr.cc/virtual/2025/poster/31022 abstract： Adaptive gradient methods have been widely adopted in training large-scale deep neural networks, especially large foundation models. Despite the huge success in practice, their theoretical advantages over classical gradient methods with uniform step sizes across all coordinates (e.g. SGD) have not been fully understood, especially in the large batch-size setting commonly used in practice. This is because the only theoretical result that can demonstrate this benefit was obtained in the original paper of Adagrad for convex nonsmooth objective functions, which is insufficient for large batch algorithms. In this work, we attempt to resolve this gap between theory and practice by proposing a novel anisotropic generalized smoothness assumption and providing corresponding analysis of Adagrad. It is shown that under anisotropic smoothness and noise conditions, AdaGrad can achieve faster convergence guarantees in terms of better dimensional dependence than algorithms with uniform step sizes across all coordinates. Experiments in logistic regression and instruction following fine-tuning tasks provide strong evidence to support our novel assumption and theoretical analysis.

1607. Re-evaluating Open-ended Evaluation of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28571 abstract： Evaluation has traditionally focused on ranking candidates for a specific skill. Modern generalist models, such as Large Language Models (LLMs), decidedly outpace this paradigm. Open-ended evaluation systems, where candidate models are compared on user-submitted prompts, have emerged as a popular solution. Despite their many advantages, we show that the current Elo-based rating systems can be susceptible to and even reinforce biases in data, intentional or accidental, due to their sensitivity to redundancies. To address this issue, we propose evaluation as a 3-player game, and introduce novel game-theoretic solution concepts to ensure robustness to redundancy. We show that our method leads to intuitive ratings and provide insights into the competitive landscape of LLM development.

1608. A Skewness-Based Criterion for Addressing Heteroscedastic Noise in Causal Discovery

链接：https://iclr.cc/virtual/2025/poster/27676 abstract： Real-world data often violates the equal-variance assumption (homoscedasticity), making it essential to account for heteroscedastic noise in causal discovery. In this work, we explore heteroscedastic symmetric noise models (HSNMs), where the effect Y is modeled as $Y = f(X) + \sigma(X)N$, with X as the cause and N as independent noise following a symmetric distribution. We introduce a novel criterion for identifying HSNMs based on the skewness of the score (i.e., the gradient of the log density) of the data distribution. This criterion establishes a computationally tractable measurement that is zero in the causal direction but nonzero in the anticausal direction, enabling the causal direction discovery. We extend this skewness-based criterion to the multivariate setting and propose \texttt{SkewScore}, an algorithm that handles heteroscedastic noise without requiring the extraction of exogenous noise. We also conduct a case study on the robustness of \texttt{SkewScore} in a bivariate model with a latent confounder, providing theoretical insights into its performance. Empirical studies further validate the effectiveness of the proposed method.

1609. Towards Fast, Specialized Machine Learning Force Fields: Distilling Foundation Models via Energy Hessians

链接：https://iclr.cc/virtual/2025/poster/31192 abstract： The foundation model (FM) paradigm is transforming Machine Learning Force Fields (MLFFs), leveraging general-purpose representations and scalable training to perform a variety of computational chemistry tasks. Although MLFF FMs have begun to close the accuracy gap relative to first-principles methods, there is still a strong need for faster inference speed. Additionally, while research is increasingly focused on general-purpose models which transfer across chemical space, practitioners typically only study a small subset of systems at a given time. At test time, MLFFs must also obey physical constraints unique to the downstream use case, such as energy conservation for molecular dynamics simulations. This underscores the need for fast, specialized MLFFs relevant to specific downstream applications, which preserve test-time physical soundness while maintaining train-time scalability. In this work, we introduce a method for transferring general-purpose representations from MLFF foundation models to smaller, faster MLFFs specialized to specific

regions of chemical space. We formulate our approach as an architecture-agnostic knowledge distillation procedure, where the smaller "student" MLFF is trained to match the Hessians of the energy predictions of the "teacher" foundation model. We demonstrate our approach across multiple recent foundation models, large-scale datasets, chemical subsets, and downstream tasks. Our specialized MLFFs can be up to 20 times faster than the original foundation model, while retaining, and in some cases exceeding, its performance and that of undistilled models. We also show that distilling from a teacher model with a direct force parameterization into a student model trained with conservative forces (i.e., computed as derivatives of the potential energy) successfully leverages the representations from the large-scale teacher for improved accuracy, while maintaining energy conservation during test-time molecular dynamics simulations. More broadly, our work suggests a new paradigm for MLFF development, in which foundation models are released along with smaller, specialized simulation ``engines" for common chemical subsets. The implementation of our method is available at https://github.com/ASK-Berkeley/MLFF-distill.

1610. Exploratory Preference Optimization: Harnessing Implicit Q*-Approximation for Sample-Efficient RLHF

链接：https://iclr.cc/virtual/2025/poster/29685 abstract： This paper investigates a basic question in reinforcement learning from human feedback (RLHF) from a theoretical perspective: how to efficiently explore in an online manner under preference feedback and general function approximation. We take the initial step towards a theoretical understanding of this problem by proposing a novel algorithm, Exploratory Preference Optimization (XPO). This algorithm is elegantly simple---requiring only a one-line modification to (online) Direct Preference Optimization (DPO; Rafailov et al., 2023)---yet provides the strongest known provable guarantees. XPO augments the DPO objective with a novel and principled exploration bonus, enabling the algorithm to strategically explore beyond the support of the initial model and preference feedback data. We prove that XPO is provably sample-efficient and converges to a near-optimal policy under natural exploration conditions, regardless of the initial model's coverage. Our analysis builds on the observation that DPO implicitly performs a form of Bellman error minimization. It synthesizes previously disparate techniques from language modeling and theoretical reinforcement learning in a serendipitous fashion through the lens of KL-regularized Markov decision processes.

1611. Beyond-Expert Performance with Limited Demonstrations: Efficient Imitation Learning with Double Exploration

链接：https://iclr.cc/virtual/2025/poster/30314 abstract： Imitation learning is a central problem in reinforcement learning where the goal is to learn a policy that mimics the expert's behavior. In practice, it is often challenging to learn the expert policy from a limited number of demonstrations accurately due to the complexity of the state space. Moreover, it is essential to explore the environment and collect data to achieve beyond-expert performance. To overcome these challenges, we propose a novel imitation learning algorithm called Imitation Learning with Double Exploration (ILDE), which implements exploration in two aspects: (1) optimistic policy optimization via an exploration bonus that rewards state-action pairs with high uncertainty to potentially improve the convergence to the expert policy, and (2) curiosity-driven exploration of the states that deviate from the demonstration trajectories to potentially yield beyond-expert performance. Empirically, we demonstrate that ILDE outperforms the state-of-the-art imitation learning algorithms in terms of sample efficiency and achieves beyond-expert performance on Atari and MuJoCo tasks with fewer demonstrations than in previous work. We also provide a theoretical justification of ILDE as an uncertainty-regularized policy optimization method with optimistic exploration, leading to a regret growing sublinearly in the number of episodes.

1612. PPT: Patch Order Do Matters In Time Series Pretext Task

链接：https://iclr.cc/virtual/2025/poster/30790 abstract： Recently, patch-based models have been widely discussed in time series analysis. However, existing pretext tasks for patch-based learning, such as masking, may not capture essential time and channel-wise patch interdependencies in time series data, presumed to result in subpar model performance. In this work, we introduce Patch order-aware Pretext Task (PPT), a new self-supervised patch order learning pretext task for time series classification. PPT exploits the intrinsic sequential order information among patches across time and channel dimensions of time series data, where model training is aided by channel-wise patch permutations. The permutation disrupts patch order consistency across time and channel dimensions with controlled intensity to provide supervisory signals for learning time series order characteristics. To this end, we propose two patch order-aware learning methods: patch order consistency learning, which quantifies patch order correctness, and contrastive learning, which distinguishes weakly permuted patch sequences from strongly permuted ones. With patch order learning, we observe enhanced model performance, e.g., improving up to 7% accuracy for the supervised cardiogram task and outperforming mask-based learning by 5% in the self-supervised human activity recognition task. We also propose ACF-CoS, an evaluation metric that measures the importance of orderness for time series datasets, which enables pre-examination of the efficacy of PPT in model training.

1613. Bridging Jensen Gap for Max-Min Group Fairness Optimization in Recommendation

链接：https://iclr.cc/virtual/2025/poster/31203 abstract： Group max-min fairness (MMF) is commonly used in fairness-aware recommender systems (RS) as an optimization objective, as it aims to protect marginalized item groups and ensures a fair competition platform. However, our theoretical analysis indicates that integrating MMF constraint violates the assumption of

sample independence during optimization, causing the loss function to deviate from linear additivity. Such nonlinearity property introduces the Jensen gap between the model's convergence point and the optimal point if mini-batch sampling is applied. Both theoretical and empirical studies show that as the mini-batch size decreases and the group size increases, the Jensen gap will widen accordingly. Some methods using heuristic re-weighting or debiasing strategies have the potential to bridge the Jensen gap. However, they either lack theoretical guarantees or suffer from heavy computational costs. To overcome these limitations, we first theoretically demonstrate that the MMF-constrained objective can be essentially reformulated as a group-weighted optimization objective. Then we present an efficient and effective algorithm named FairDual, which utilizes a dual optimization technique to minimize Jensen gap. Our theoretical analysis demonstrates that FairDual can achieve a sub-linear convergence rate to the globally optimal solution and the Jensen gap can be well bounded under a mini-batch sampling strategy with random shuffle. Extensive experiments conducted using six large-scale RS backbone models on three publicly available datasets demonstrate that FairDual outperforms all baselines in terms of both accuracy and fairness.

1614. Efficient Discovery of Pareto Front for Multi-Objective Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28888 abstract： Multi-objective reinforcement learning (MORL) excels at handling rapidly changing preferences in tasks that involve multiple criteria, even for unseen preferences. However, previous dominating MORL methods typically generate a fixed policy set or preference-conditioned policy through multiple training iterations exclusively for sampled preference vectors, and cannot ensure the efficient discovery of the Pareto front. Furthermore, integrating preferences into the input of policy or value functions presents scalability challenges, in particular as the dimension of the state and preference space grow, which can complicate the learning process and hinder the algorithm's performance on more complex tasks. To address these issues, we propose a two-stage Pareto front discovery algorithm called Constrained MORL (C-MORL), which serves as a seamless bridge between constrained policy optimization and MORL. Concretely, a set of policies is trained in parallel in the initialization stage, with each optimized towards its individual preference over the multiple objectives. Then, to fill the remaining vacancies in the Pareto front, the constrained optimization steps are employed to maximize one objective while constraining the other objectives to exceed a predefined threshold. Empirically, compared to recent advancements in MORL methods, our algorithm achieves more consistent and superior performances in terms of hypervolume, expected utility, and sparsity on both discrete and continuous control tasks, especially with numerous objectives (up to nine objectives in our experiments).

1615. Constructing Confidence Intervals for Average Treatment Effects from Multiple Datasets

链接：https://iclr.cc/virtual/2025/poster/30580 abstract： Constructing confidence intervals (CIs) for the average treatment effect (ATE) from patient records is crucial to assess the effectiveness and safety of drugs. However, patient records typically come from different hospitals, thus raising the question of how multiple observational/experimental datasets can be effectively combined for this purpose. In our paper, we propose a new method that estimates the ATE from multiple observational/experimental datasets and provides valid CIs. Our method makes little assumptions about the observational datasets and is thus widely applicable in medical practice. The key idea of our method is that we leverage prediction-powered inferences and thereby essentially `shrink' the CIs so that we offer more precise uncertainty quantification as compared to na{\"i}ve approaches. We further prove the unbiasedness of our method and the validity of our CIs. We confirm our theoretical results through various numerical experiments.

1616. Efficient Training of Neural Stochastic Differential Equations by Matching Finite Dimensional Distributions

链接：https://iclr.cc/virtual/2025/poster/29011 abstract： Neural Stochastic Differential Equations (Neural SDEs) have emerged as powerful mesh-free generative models for continuous stochastic processes, with critical applications in fields such as finance, physics, and biology. Previous state-of-the-art methods have relied on adversarial training, such as GANs, or on minimizing distance measures between processes using signature kernels. However, GANs suffer from issues like instability, mode collapse, and the need for specialized training techniques, while signature kernel-based methods require solving linear PDEs and backpropagating gradients through the solver, whose computational complexity scales quadratically with the discretization steps. In this paper, we identify a novel class of strictly proper scoring rules for comparing continuous Markov processes. This theoretical finding naturally leads to a novel approach called Finite Dimensional Matching (FDM) for training Neural SDEs. Our method leverages the Markov property of SDEs to provide a computationally efficient training objective. This scoring rule allows us to bypass the computational overhead associated with signature kernels and reduces the training complexity from $O(D^2)$ to $O(D)$ per epoch, where D represents the number of discretization steps of the process. We demonstrate that FDM achieves superior performance, consistently outperforming existing methods in terms of both computational efficiency and generative quality.

1617. The Computational Complexity of Circuit Discovery for Inner Interpretability

链接：https://iclr.cc/virtual/2025/poster/29673 abstract： Many proposed applications of neural networks in machine learning,

cognitive/brain science, and society hinge on the feasibility of inner interpretability via circuit discovery. This calls for empirical and theoretical explorations of viable algorithmic options. Despite advances in the design and testing of heuristics, there are concerns about their scalability and faithfulness at a time when we lack understanding of the complexity properties of the problems they are deployed to solve. To address this, we study circuit discovery with classical and parameterized computational complexity theory: (1) we describe a conceptual scaffolding to reason about circuit finding queries in terms of affordances for description, explanation, prediction and control; (2) we formalize a comprehensive set of queries for mechanistic explanation, and propose a formal framework for their analysis; (3) we use it to settle the complexity of many query variants and relaxations of practical interest on multi-layer perceptrons. Our findings reveal a challenging complexity landscape. Many queries are intractable, remain fixed-parameter intractable relative to model/circuit features, and inapproximable under additive, multiplicative, and probabilistic approximation schemes. To navigate this landscape, we prove there exist transformations to tackle some of these hard problems with better-understood heuristics, and prove the tractability or fixed-parameter tractability of more modest queries which retain useful affordances. This framework allows us to understand the scope and limits of interpretability queries, explore viable options, and compare their resource demands on existing and future architectures.

1618. ImpScore: A Learnable Metric For Quantifying The Implicitness Level of Sentences

链接：https://iclr.cc/virtual/2025/poster/28811 abstract： Handling implicit language is essential for natural language processing systems to achieve precise text understanding and facilitate natural interactions with users. Despite its importance, the absence of a metric for accurately measuring the implicitness of language significantly constrains the depth of analysis possible in evaluating models' comprehension capabilities. This paper addresses this gap by developing a scalar metric that quantifies the implicitness level of language without relying on external references. Drawing on principles from traditional linguistics, we define "implicitness" as the divergence between semantic meaning and pragmatic interpretation. To operationalize this definition, we introduce ImpScore, a reference-free metric formulated through an interpretable regression model. This model is trained using pairwise contrastive learning on a specially curated dataset consisting of (implicit sentence, explicit sentence) pairs. We validate ImpScore through a user study that compares its assessments with human evaluations on out-of-distribution data, demonstrating its accuracy and strong correlation with human judgments. Additionally, we apply ImpScore to hate speech detection datasets, illustrating its utility and highlighting significant limitations in current large language models' ability to understand highly implicit content. Our metric is publicly available at https://github.com/audreycs/ImpScore.

1619. Immunogenicity Prediction with Dual Attention Enables Vaccine Target Selection

链接：https://iclr.cc/virtual/2025/poster/28756 abstract： Immunogenicity prediction is a central topic in reverse vaccinology for finding candidate vaccines that can trigger protective immune responses. Existing approaches typically rely on highly compressed features and simple model architectures, leading to limited prediction accuracy and poor generalizability. To address these challenges, we introduce VenusVaccine, a novel deep learning solution with a dual attention mechanism that integrates pre-trained latent vector representations of protein sequences and structures. We also compile the most comprehensive immunogenicity dataset to date, encompassing over 7000 antigen sequences, structures, and immunogenicity labels from bacteria, viruses, and tumors. Extensive experiments demonstrate that VenusVaccine outperforms existing methods across a wide range of evaluation metrics. Furthermore, we establish a post-hoc validation protocol to assess the practical significance of deep learning models in tackling vaccine design challenges. Our work provides an effective tool for vaccine design and sets valuable benchmarks for future research. The implementation is at \url{https://github.com/songleee/VenusVaccine}.

1620. ConvCodeWorld: Benchmarking Conversational Code Generation in Reproducible Feedback Environments

链接：https://iclr.cc/virtual/2025/poster/28169 abstract： Large language models (LLMs) have proven invaluable for code generation, particularly in interactive settings. However, existing code generation benchmarks fail to capture the diverse feedback encountered in multi-turn interactions, limiting our ability to evaluate LLMs in these contexts. To address this gap, we present a set of novel benchmarks that explicitly model the quality of feedback provided to code generation LLMs. Our contributions are threefold: First, we introduce CONVCODEWORLD, a novel and reproducible environment for benchmarking interactive code generation. CONVCODEWORLD simulates 9 distinct interactive code generation scenarios while systematically combining three types of feedback: (a) compilation feedback; (b) execution feedback with varying test coverage; (c) verbal feedback generated by GPT-4o with different levels of expertise. Second, we introduce CONVCODEBENCH, a fast, static version of benchmark that uses pre-generated feedback logs, eliminating the need for costly dynamic verbal feedback generation while maintainingstrong Spearman's rank correlations (0.82 to 0.99) with CONVCODEWORLD. Third, extensive evaluations of both closed-source and open-source LLMs including R1-Distill on CONVCODEWORLD reveal key insights: (a) LLM performance varies significantly based on the feedback provided; (b) Weaker LLMs, with sufficient feedback, can outperform single-turn results of state-of-the-art LLMs without feedback; (c) Training on a specific feedback combination can limit an LLM's ability to utilize unseen combinations; (d) LLMs solve problems in fewer turns (high MRR) may not solve as many problems overall (high Recall), and vice versa. All implementations and benchmarks will be made publicly available at https://huggingface.co/spaces/ConvCodeWorld/ConvCodeWorld

1621. Counterfactual Concept Bottleneck Models

链接：https://iclr.cc/virtual/2025/poster/27870 abstract： Current deep learning models are not designed to simultaneously address three fundamental questions: predict class labels to solve a given classification task (the "What?"), simulate changes in the situation to evaluate how this impacts class predictions (the "How?"), and imagine how the scenario should change to result in different class predictions (the "Why not?"). While current approaches in causal representation learning and concept interpretability are designed to address some of these questions individually (such as Concept Bottleneck Models, which address both what" andhow" questions), no current deep learning model is specifically built to answer all of them at the same time. To bridge this gap, we introduce CounterFactual Concept Bottleneck Models (CF-CBMs), a class of models designed to efficiently address the above queries all at once without the need to run post-hoc searches. Our experimental results demonstrate that CF-CBMs: achieve classification accuracy comparable to black-box models and existing CBMs (“What?”), rely on fewer important concepts leading to simpler explanations (“How?”), and produce interpretable, concept-based counterfactuals (“Why not?”). Additionally, we show that training the counterfactual generator jointly with the CBM leads to two key improvements: (i) it alters the model's decision-making process, making the model rely on fewer important concepts (leading to simpler explanations), and (ii) it significantly increases the causal effect of concept interventions on class predictions, making the model more responsive to these changes.

1622. “I Am the One and Only, Your Cyber BFF”: Understanding the Impact of GenAI Requires Understanding the Impact of Anthropomorphic AI

链接：https://iclr.cc/virtual/2025/poster/31358 abstract： Many state-of-the-art generative AI (GenAI) systems are increasingly prone to anthropomorphic behaviors, i.e., to generating outputs that are perceived to be human-like. While this has led to scholars increasingly raising concerns about possible negative impacts such anthropomorphic AI systems can give rise to, anthropomorphism in AI development, deployment, and use remains vastly overlooked, understudied, and underspecified. In this blog post, we argue that we cannot thoroughly map the social impacts of generative AI without mapping the social impacts of anthropomorphic AI, and outline a call to action.

1623. Predicate Hierarchies Improve Few-Shot State Classification

链接：https://iclr.cc/virtual/2025/poster/28492 abstract： State classification of objects and their relations is core to many long-horizon tasks, particularly in robot planning and manipulation. However, the combinatorial explosion of possible object-predicate combinations, coupled with the need to adapt to novel real-world environments, makes it a desideratum for state classification models to generalize to novel queries with few examples. To this end, we propose PHIER, which leverages predicate hierarchies to generalize effectively in few-shot scenarios. PHIER uses an object-centric scene encoder, self-supervised losses that infer semantic relations between predicates, and a hyperbolic distance metric that captures hierarchical structure; it learns a structured latent space of image-predicate pairs that guides reasoning over state classification queries. We evaluate PHIER in the CALVIN and BEHAVIOR robotic environments and show that PHIER significantly outperforms existing methods in few-shot, out-of-distribution state classification, and demonstrates strong zero- and few-shot generalization from simulated to real-world tasks. Our results demonstrate that leveraging predicate hierarchies improves performance on state classification tasks with limited data.

1624. Expand and Compress: Exploring Tuning Principles for Continual Spatio-Temporal Graph Forecasting

链接：https://iclr.cc/virtual/2025/poster/30341 abstract： The widespread deployment of sensing devices leads to a surge in data for spatio-temporal forecasting applications such as traffic flow, air quality, and wind energy. Although spatio-temporal graph neural networks (STGNNs) have achieved success in modeling various static spatio-temporal forecasting scenarios, real-world spatio-temporal data are typically received in a streaming manner, and the network continuously expands with the installation of new sensors. Thus, spatio-temporal forecasting in streaming scenarios faces dual challenges: the inefficiency of retraining models over newly-arrived data and the detrimental effects of catastrophic forgetting over long-term history. To address these challenges, we propose a novel prompt tuning-based continuous forecasting method, EAC, following two fundamental tuning principles guided by empirical and theoretical analysis: expand and compress, which effectively resolve the aforementioned problems with lightweight tuning parameters. Specifically, we integrate the base STGNN with a continuous prompt pool, utilizing stored prompts (\ie, few learnable parameters) in memory, and jointly optimize them with the base STGNN. This method ensures that the model sequentially learns from the spatio-temporal data stream to accomplish tasks for corresponding periods. Extensive experimental results on multiple real-world datasets demonstrate the multi-faceted superiority of EAC over the state-of-the-art baselines, including effectiveness, efficiency, universality, etc.

1625. Air Quality Prediction with Physics-Guided Dual Neural ODEs in Open Systems

链接：https://iclr.cc/virtual/2025/poster/28589 abstract： Air pollution significantly threatens human health and ecosystems, necessitating effective air quality prediction to inform public policy. Traditional approaches are generally categorized into physics-based and data-driven models. Physics-based models usually struggle with high computational demands and closed-

system assumptions, while data-driven models may overlook essential physical dynamics, confusing the capturing of spatiotemporal correlations. Although some physics-guided approaches combine the strengths of both models, they often face a mismatch between explicit physical equations and implicit learned representations. To address these challenges, we propose Air-DualODE, a novel physics-guided approach that integrates dual branches of Neural ODEs for air quality prediction. The first branch applies open-system physical equations to capture spatiotemporal dependencies for learning physics dynamics, while the second branch identifies the dependencies not addressed by the first in a fully data-driven way. These dual representations are temporally aligned and fused to enhance prediction accuracy. Our experimental results demonstrate that Air-DualODE achieves state-of-the-art performance in predicting pollutant concentrations across various spatial scales, thereby offering a promising solution for real-world air quality challenges.

1626. Track-On: Transformer-based Online Point Tracking with Memory

链接：https://iclr.cc/virtual/2025/poster/28358 abstract：

1627. Towards Neural Scaling Laws for Time Series Foundation Models

链接：https://iclr.cc/virtual/2025/poster/27992 abstract：

1628. Enhancing Document Understanding with Group Position Embedding: A Novel Approach to Incorporate Layout Information

链接：https://iclr.cc/virtual/2025/poster/30445 abstract：

1629. First-Person Fairness in Chatbots

链接：https://iclr.cc/virtual/2025/poster/29520 abstract： Evaluating chatbot fairness is crucial given their rapid proliferation, yet typical chatbot tasks (e.g., resume writing, entertainment) diverge from the institutional decision-making tasks (e.g., resume screening) which have traditionally been central to discussion of algorithmic fairness. The open-ended nature and diverse use-cases of chatbots necessitate novel methods for bias assessment. This paper addresses these challenges by introducing a scalable counterfactual approach to evaluate "first-person fairness," meaning fairness toward chatbot users based on demographic characteristics. Our method employs a Language Model as a Research Assistant (LMRA) to yield quantitative measures of harmful stereotypes and qualitative analyses of demographic differences in chatbot responses. We apply this approach to assess biases in six of our language models across millions of interactions, covering sixty-six tasks in nine domains and spanning two genders and four races. Independent human annotations corroborate the LMRA-generated bias evaluations. This study represents the first large-scale fairness evaluation based on real-world chat data. We highlight that post-training reinforcement learning techniques significantly mitigate these biases. This evaluation provides a practical methodology for ongoing bias monitoring and mitigation.

1630. Interactive Adjustment for Human Trajectory Prediction with Individual Feedback

链接：https://iclr.cc/virtual/2025/poster/30472 abstract： Human trajectory prediction is fundamental for autonomous driving and service robot. The research community has studied various important aspects of this task and made remarkable progress recently. However, there is an essential perspective which is not well exploited in previous research all along, namely individual feedback. Individual feedback exists in the sequential nature of trajectory prediction, where earlier predictions of a target can be verified over time by his ground-truth trajectories to obtain feedback which provides valuable experience for subsequent predictions on the same agent. In this paper, we show such feedback can reveal the strengths and weaknesses of the model's predictions on a specific target and heuristically guide to deliver better predictions on him. We present an interactive adjustment network to effectively model and leverage the feedback. This network first exploits the feedback from previous predictions to dynamically generate an adjuster which then interactively makes appropriate adjustments to current predictions for more accurate ones. We raise a novel displacement expectation loss to train this interactive architecture. Through experiments on representative prediction methods and widely-used benchmarks, we demonstrate the great value of individual feedback and the superior effectiveness of proposed interactive adjustment network.

1631. On the Modeling Capabilities of Large Language Models for Sequential Decision Making

链接：https://iclr.cc/virtual/2025/poster/27887 abstract： Large pretrained models are showing increasingly better performance in reasoning and planning tasks across different modalities, opening the possibility to leverage them for complex sequential decision making problems. In this paper, we investigate the capabilities of Large Language Models (LLMs) for reinforcement learning (RL) across a diversity of interactive domains. We evaluate their ability to produce decision-making policies, either directly, by generating actions, or indirectly, by first generating reward models to train an agent with RL. Our results show that, even without task-specific fine-tuning, LLMs excel at reward modeling. In particular, crafting rewards through artificial intelligence

(AI) feedback yields the most generally applicable approach and can enhance performance by improving credit assignment and exploration. Finally, in environments with unfamiliar dynamics, we explore how fine-tuning LLMs with synthetic data can significantly improve their reward modeling capabilities while mitigating catastrophic forgetting, further broadening their utility in sequential decision-making tasks.

1632. AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution

链接：https://iclr.cc/virtual/2025/poster/30669 abstract： The influence of contextual input on the behavior of large language models (LLMs) has prompted the development of context attribution methods that aim to quantify each context span's effect on an LLM's generations. The leave-one-out (LOO) error, which measures the change in the likelihood of the LLM's response when a given span of the context is removed, provides a principled way to perform context attribution, but can be prohibitively expensive to compute for large models. In this work, we introduce AttriBoT, a series of novel techniques for efficiently computing an approximation of the LOO error for context attribution. Specifically, AttriBoT uses cached activations to avoid redundant operations, performs hierarchical attribution to reduce computation, and emulates the behavior of large target models with smaller proxy models. Taken together, AttriBoT can provide a 300x speedup while remaining more faithful to a target model's LOO error than prior context attribution methods. This stark increase in performance makes computing context attributions for a given response $30\times$ faster than generating the response itself, empowering real-world applications that require computing attributions at scale. We release a user-friendly and efficient implementation of AttriBoT to enable efficient LLM interpretability as well as encourage future development of efficient context attribution methods.

1633. LeanQuant: Accurate and Scalable Large Language Model Quantization with Loss-error-aware Grid

链接：https://iclr.cc/virtual/2025/poster/30168 abstract： Large language models (LLMs) have shown immense potential across various domains, but their high memory requirements and inference costs remain critical challenges for deployment. Post-training quantization (PTQ) has emerged as a promising technique to reduce memory requirements and decoding latency. However, recent accurate quantization methods often depend on specialized computations or custom data formats to achieve better model quality, which limits their compatibility with popular frameworks, as they require dedicated inference kernels tailored to specific hardware and software platforms, hindering wider adoption. Furthermore, many competitive methods have high resource requirements and computational overhead for quantizing models, making it challenging to scale them to hundreds of billions of parameters. In response to these challenges, we propose LeanQuant (Loss-error-aware network Quantization), a novel quantization method that is accurate, versatile, and scalable. In the existing popular iterative loss-error-based quantization framework, we identify a critical limitation in prior methods: the min-max affine quantization grid fails to preserve model quality due to outliers in inverse Hessian diagonals. To overcome this fundamental issue, we propose learning loss-error-aware grids, instead of using non-adaptive min-max affine grids. Our approach not only produces quantized models that are more accurate but also generalizes to a wider range of quantization types, including affine and non-uniform quantization, enhancing compatibility with more frameworks. Extensive experiments with recent LLMs demonstrate that LeanQuant is highly accurate, comparing favorably against competitive baselines in model quality, and scalable, achieving very accurate quantization of Llama-3.1 405B, one of the largest open-source LLMs to date, using two Quadro RTX 8000-48GB GPUs in 21 hours. Our code is available at https://github.com/LeanModels/LeanQuant.

1634. PIG: Physics-Informed Gaussians as Adaptive Parametric Mesh Representations

链接：https://iclr.cc/virtual/2025/poster/27752 abstract： The numerical approximation of partial differential equations (PDEs) using neural networks has seen significant advancements through Physics-Informed Neural Networks (PINNs). Despite their straightforward optimization framework and flexibility in implementing various PDEs, PINNs often suffer from limited accuracy due to the spectral bias of Multi-Layer Perceptrons (MLPs), which struggle to effectively learn high-frequency and nonlinear components. Recently, parametric mesh representations in combination with neural networks have been investigated as a promising approach to eliminate the inductive bias of MLPs. However, they usually require high-resolution grids and a large number of collocation points to achieve high accuracy while avoiding overfitting. In addition, the fixed positions of the mesh parameters restrict their flexibility, making accurate approximation of complex PDEs challenging. To overcome these limitations, we propose Physics-Informed Gaussians (PIGs), which combine feature embeddings using Gaussian functions with a lightweight neural network. Our approach uses trainable parameters for the mean and variance of each Gaussian, allowing for dynamic adjustment of their positions and shapes during training. This adaptability enables our model to optimally approximate PDE solutions, unlike models with fixed parameter positions. Furthermore, the proposed approach maintains the same optimization framework used in PINNs, allowing us to benefit from their excellent properties. Experimental results show the competitive performance of our model across various PDEs, demonstrating its potential as a robust tool for solving complex PDEs. Our project page is available at https://namgyukang.github.io/Physics-Informed-Gaussians/

1635. Transformer Encoder Satisfiability: Complexity and Impact on Formal Reasoning

链接：https://iclr.cc/virtual/2025/poster/29412 abstract： We analyse the complexity of the satisfiability problem, or similarly feasibility problem, (trSAT) for transformer encoders (TE), which naturally occurs in formal verification or interpretation, collectively referred to as formal reasoning. We find that trSAT is undecidable when considering TE as they are commonly studied in the expressiveness community. Furthermore, we identify practical scenarios where trSAT is decidable and establish corresponding complexity bounds. Beyond trivial cases, we find that quantized TE, those restricted by fixed-width arithmetic, lead to the decidability of trSAT due to their limited attention capabilities. However, the problem remains difficult, as we establish scenarios where trSAT is NEXPTIME-hard and others where it is solvable in NEXPTIME for quantized TE. To complement our complexity results, we place our findings and their implications in the broader context of formal reasoning.

1636. Learning Causal Alignment for Reliable Disease Diagnosis

链接：https://iclr.cc/virtual/2025/poster/28327 abstract： Aligning the decision-making process of machine learning algorithms with that of experienced radiologists is crucial for reliable diagnosis. While existing methods have attempted to align their prediction behaviors to those of radiologists reflected in the training data, this alignment is primarily associational rather than causal, resulting in pseudo-correlations that may not transfer well. In this paper, we propose a causality-based alignment framework towards aligning the model's decision process with that of experts. Specifically, we first employ counterfactual generation to identify the causal chain of model decisions. To align this causal chain with that of experts, we propose a causal alignment loss that enforces the model to focus on causal factors underlying each decision step in the whole causal chain. To optimize this loss that involves the counterfactual generator as an implicit function of the model's parameters, we employ the implicit function theorem equipped with the conjugate gradient method for efficient estimation. We demonstrate the effectiveness of our method on two medical diagnosis applications, showcasing faithful alignment to radiologists.

1637. Refining CLIP's Spatial Awareness: A Visual-Centric Perspective

链接：https://iclr.cc/virtual/2025/poster/31093 abstract： Contrastive Language-Image Pre-training (CLIP) excels in global alignment with language but exhibits limited sensitivity to spatial information, leading to strong performance in zero-shot classification tasks but underperformance in tasks requiring precise spatial understanding. Recent approaches have introduced Region-Language Alignment (RLA) to enhance CLIP's performance in dense multimodal tasks by aligning regional visual representations with corresponding text inputs. However, we find that CLIP ViTs fine-tuned with RLA suffer from notable loss in spatial awareness, which is crucial for dense prediction tasks. To address this, we propose the Spatial Correlation Distillation (SCD) framework, which preserves CLIP's inherent spatial structure and mitigates above degradation. To further enhance spatial correlations, we introduce a lightweight Refiner that extracts refined correlations directly from CLIP before feeding them into SCD, based on an intriguring finding that CLIP naturally capture high-quality dense features. Together, these components form a robust distillation framework that enables CLIP ViTs to integrate both visual-language and visual-centric improvements, achieving state-of-the-art results across various open-vocabulary dense prediction benchmarks.

1638. Near-Optimal Online Learning for Multi-Agent Submodular Coordination: Tight Approximation and Communication Efficiency

链接：https://iclr.cc/virtual/2025/poster/28714 abstract： Coordinating multiple agents to collaboratively maximize submodular functions in unpredictable environments is a critical task with numerous applications in machine learning, robot planning and control. The existing approaches, such as the OSG algorithm, are often hindered by their poor approximation guarantees and the rigid requirement for a fully connected communication graph. To address these challenges, we firstly present a $\textbf{MA-OSMA}$ algorithm, which employs the multi-linear extension to transfer the discrete submodular maximization problem into a continuous optimization, thereby allowing us to reduce the strict dependence on a complete graph through consensus techniques. Moreover, $\textbf{MA-OSMA}$ leverages a novel surrogate gradient to avoid sub-optimal stationary points. To eliminate the computationally intensive projection operations in $\textbf{MA-OSMA}$, we also introduce a projection-free $\textbf{MA-OSEA}$ algorithm, which effectively utilizes the KL divergence by mixing a uniform distribution. Theoretically, we confirm that both algorithms achieve a regret bound of $\widetilde{O}(\sqrt{\frac{C_{T}T}{1-\beta}})$ against a $(\frac{1-e^{-c}}{c})$-approximation to the best comparator in hindsight, where C_{T} is the deviation of maximizer sequence, β is the spectral gap of the network and c is the joint curvature of submodular objectives. This result significantly improves the $(\frac{1}{1+c})$-approximation provided by the state-of-the-art OSG algorithm. Finally, we demonstrate the effectiveness of our proposed algorithms through simulation-based multi-target tracking.

1639. YouTube-SL-25: A Large-Scale, Open-Domain Multilingual Sign Language Parallel Corpus

链接：https://iclr.cc/virtual/2025/poster/28414 abstract：

1640. AIR-BENCH 2024: A Safety Benchmark based on Regulation and Policies Specified Risk Categories

链接：https://iclr.cc/virtual/2025/poster/29470 abstract：

1641. Computational Limits of Low-Rank Adaptation (LoRA) Fine-Tuning for Transformer Models

链接：https://iclr.cc/virtual/2025/poster/29984 abstract： We study the computational limits of Low-Rank Adaptation (LoRA) for finetuning transformer-based models using fine-grained complexity theory.Our key observation is that the existence of low-rank decompositions within the gradient computation of LoRA adaptation leads to possible algorithmic speedup.This allows us to (i) identify a phase transition behavior of efficiency \blue{assuming the Strong Exponential Time Hypothesis (SETH)}, and (ii) prove the existence of almost linear algorithms by controlling the LoRA update computation term by term.For the former, we identify a sharp transition in the efficiency of all possible rank-r LoRA update algorithms for transformers, based on specific norms resulting from the multiplications of the input sequence X, pretrained weights ${W^\star}$, and adapter matrices $\alpha B A/r$.Specifically, we derive a shared upper bound threshold for such norms and show that efficient (sub-quadratic) approximation algorithms of LoRA exist only below this threshold.For the latter, we prove the existence of almost linear approximation algorithms for LoRA adaptation by utilizing the hierarchical low-rank structures of LoRA gradients and approximating the gradients with a series of chained low-rank approximations.To showcase our theory, we consider two practical scenarios: partial (e.g., only W_V and W_Q) and full adaptations (e.g., W_Q, W_V, and W_K) of weights in attention heads.

1642. Mini-batch Coresets for Memory-efficient Language Model Training on Data Mixtures

链接：https://iclr.cc/virtual/2025/poster/29123 abstract： Training with larger mini-batches improves the convergence rate and can yield superior performance. However, training with large mini-batches becomes prohibitive for Large Language Models (LLMs), due to the large GPU memory requirement. To address this problem, an effective approach is finding small mini-batch coresets that closely match the gradient of larger mini-batches. However, this approach becomes infeasible and ineffective for LLMs, due to the highly imbalanced mixture of sources in language data, use of the Adam optimizer, and the very large gradient dimensionality of LLMs. In this work, we address the above challenges by proposing Coresets for Training LLMs (CoLM). First, we show that mini-batch coresets found by gradient matching do not contain representative examples of the small sources w.h.p., and thus including all examples of the small sources in the mini-batch coresets is crucial for optimal performance. Second, we normalize the gradients by their historical exponential to find mini-batch coresets for training with Adam. Finally, we leverage zeroth-order methods to find smooth gradient of the last V-projection matrix and sparsify it to keep the dimensions with the largest normalized gradient magnitude. We apply CoLM to fine-tuning Phi-2, Phi-3, Zephyr, and Llama-3 models with LoRA on MathInstruct and SuperGLUE benchmark. Remarkably, CoLM reduces the memory requirement of fine-tuning by 2x and even outperforms training with 4x larger mini-batches. Moreover, CoLM seamlessly integrates with existing memory-efficient training methods like LoRA, further reducing the memory requirements of training LLMs.

1643. QA-Calibration of Language Model Confidence Scores

链接：https://iclr.cc/virtual/2025/poster/30480 abstract： To use generative question-and-answering (QA) systems for decision-making and in any critical application, these systems need to provide well-calibrated confidence scores that reflect the correctness of their answers. Existing calibration methods aim to ensure that the confidence score is, on average, indicative of the likelihood that the answer is correct. We argue, however, that this standard (average-case) notion of calibration is difficult to interpret for decision-making in generative QA. To address this, we generalize the standard notion of average calibration and introduce QA-calibration, which ensures calibration holds across different question-and-answer groups. We then propose discretized posthoc calibration schemes for achieving QA-calibration. We establish distribution-free guarantees on the performance of this method and validate our method on confidence scores returned by elicitation prompts across multiple QA benchmarks and large language models (LLMs).

1644. Speech Robust Bench: A Robustness Benchmark For Speech Recognition

链接：https://iclr.cc/virtual/2025/poster/30483 abstract： As Automatic Speech Recognition (ASR) models become ever more pervasive, it is important to ensure that they make reliable predictions under corruptions present in the physical and digital world. We propose Speech Robust Bench (SRB), a comprehensive benchmark for evaluating the robustness of ASR models to diverse corruptions. SRB is composed of 114 input perturbations which simulate an heterogeneous range of corruptions that ASR models may encounter when deployed in the wild. We use SRB to evaluate the robustness of several state-of-the-art ASR models and observe that model size and certain modeling choices such as the use of discrete representations, or self-training appear to be conducive to robustness. We extend this analysis to measure the robustness of ASR models on data from various demographic subgroups, namely English and Spanish speakers, and males and females. Our results revealed noticeable disparities in the model's robustness across subgroups. We believe that SRB will significantly facilitate future research towards robust ASR models, by making it easier to conduct comprehensive and comparable robustness evaluations.

1645. Segment Any 3D Object with Language

链接：https://iclr.cc/virtual/2025/poster/30409 abstract： In this paper, we investigate Open-Vocabulary 3D Instance

Segmentation (OV-3DIS) with free-form language instructions. Earlier works mainly rely on annotated base categories for training which leads to limited generalization to unseen novel categories. To mitigate the poor generalizability to novel categories, recent works generate class-agnostic masks or projecting generalized masks from 2D to 3D, subsequently classifying them with the assistance of 2D foundation model. However, these works often disregard semantic information in the mask generation, leading to sub-optimal performance. Instead, generating generalizable but semantic-aware masks directly from 3D point clouds would result in superior outcomes. To the end, we introduce Segment any 3D Object with LanguagE ($\textbf{SOLE}$), which is a semantic and geometric-aware visual-language learning framework with strong generalizability by generating semantic-related masks directly from 3D point clouds. Specifically, we propose a multimodal fusion network to incorporate multimodal semantics in both backbone and decoder. In addition, to align the 3D segmentation model with various language instructions and enhance the mask quality, we introduce three types of multimodal associations as supervision. Our SOLE outperforms previous methods by a large margin on ScanNetv2, ScanNet200, and Replica benchmarks, and the results are even closed to the fully-supervised counterpart despite the absence of class annotations in the training. Furthermore, extensive qualitative results demonstrate the versatility of our SOLE to language instructions. The code will be made publicly available.

1646. ComPC: Completing a 3D Point Cloud with 2D Diffusion Priors

链接：https://iclr.cc/virtual/2025/poster/29579 abstract： 3D point clouds directly collected from objects through sensors are often incomplete due to self-occlusion. Conventional methods for completing these partial point clouds rely on manually organized training sets and are usually limited to object categories seen during training. In this work, we propose a test-time framework for completing partial point clouds across unseen categories without any requirement for training. Leveraging point rendering via Gaussian Splatting, we develop techniques of Partial Gaussian Initialization, Zero-shot Fractal Completion, and Point Cloud Extraction that utilize priors from pre-trained 2D diffusion models to infer missing regions and extract uniform completed point clouds. Experimental results on both synthetic and real-world scanned point clouds demonstrate that our approach outperforms existing methods in completing a variety of objects. Our project page is at \url{https://tianxinhuang.github.io/projects/ComPC/}.

1647. Beyond Autoregression: Discrete Diffusion for Complex Reasoning and Planning

链接：https://iclr.cc/virtual/2025/poster/29876 abstract： Autoregressive language models, despite their impressive capabilities, struggle with complex reasoning and long-term planning tasks. We introduce discrete diffusion models as a novel solution to these challenges. Through the lens of subgoal imbalance, we demonstrate how diffusion models effectively learn difficult subgoals that elude autoregressive approaches. We propose Multi-Granularity Diffusion Modeling (MGDM), which prioritizes subgoals based on difficulty during learning. On complex tasks like Countdown, Sudoku, and Boolean Satisfiability Problems, MGDM significantly outperforms autoregressive models without using search techniques. For instance, MGDM achieves 91.5\% and 100\% accuracy on Countdown and Sudoku, respectively, compared to 45.8\% and 20.7\% for autoregressive models. Our work highlights the potential of diffusion-based approaches in advancing AI capabilities for sophisticated language understanding and problem-solving tasks. All associated codes are available at \href{https://github.com/HKUNLP/diffusion-vs-ar}{https://github.com/HKUNLP/diffusion-vs-ar}.

1648. Beyond single neurons: population response geometry in digital twins of mouse visual cortex

链接：https://iclr.cc/virtual/2025/poster/28582 abstract： Hierarchical visual processing is essential for cognitive functions like object recognition and spatial localization. Traditional studies of the neural basis of these computations have focused on single-neuron activity, but recent advances in large-scale neural recordings emphasize the growing need to understand computations at the population level. Digital twins-computational models trained on neural data-have successfully replicated single-neuron behavior, but their effectiveness in capturing the joint activity of neurons remains unclear. In this study, we investigate how well digital twins describe population responses in mouse visual cortex. We show that these models fail to accurately represent the geometry of population activity, particularly its differentiability and how this geometry evolves across the visual hierarchy. To address this, we explore how dataset, network architecture, loss function, and training method affect the ability of digital twins to recapitulate population properties. We demonstrate that improving model alignment with experiments requires training strategies that enhance robustness and generalization, reflecting principles observed in biological systems. These findings underscore the need to evaluate digital twins from multiple perspectives, identify key areas for refinement, and establish a foundation for using these models to explore neural computations at the population level.

1649. A Conditional Independence Test in the Presence of Discretization

链接：https://iclr.cc/virtual/2025/poster/28799 abstract： Testing conditional independence (CI) has many important applications, such as Bayesian network learning and causal discovery. Although several approaches have been developed for learning CI structures for observed variables, those existing methods generally fail to work when the variables of interest can not be directly observed and only discretized values of those variables are available. For example, if X_1, $\tilde{X}_2$ and X_3 are the observed variables, where $\tilde{X}_2$ is a discretization of the latent variable X_2, applying the existing methods to the observations of X_1, $\tilde{X}_2$ and X_3 would lead to a false conclusion about the underlying CI of

variables X_1, X_2 and X_3.Motivated by this, we propose a CI test specifically designed to accommodate the presence of discretization. To achieve this, a bridge equation and nodewise regression are used to recover the precision coefficients reflecting the conditional dependence of the latent continuous variables under the nonparanormal model. An appropriate test statistic has been proposed, and its asymptotic distribution under the null hypothesis of CI has been derived.Theoretical analysis, along with empirical validation on various datasets, rigorously demonstrates the effectiveness of our testing methods.

1650. Towards Unbiased Learning in Semi-Supervised Semantic Segmentation

链接：https://iclr.cc/virtual/2025/poster/30785 abstract： Semi-supervised semantic segmentation aims to learn from a limited amount of labeled data and a large volume of unlabeled data, which has witnessed impressive progress with the recent advancement of deep neural networks. However, existing methods tend to neglect the fact of class imbalance issues, leading to the Matthew effect, that is, the poorly calibrated model's predictions can be biased towards the majority classes and away from minority classes with fewer samples. In this work, we analyze the Matthew effect present in previous methods that hinder model learning from a discriminative perspective. In light of this background, we integrate generative models into semi-supervised learning, taking advantage of their better class-imbalance tolerance. To this end, we propose DiffMatch to formulate the semi-supervised semantic segmentation task as a conditional discrete data generation problem to alleviate the Matthew effect of discriminative solutions from a generative perspective. Plus, to further reduce the risk of overfitting to the head classes and to increase coverage of the tail class distribution, we mathematically derive a debiased adjustment to adjust the conditional reverse probability towards unbiased predictions during each sampling step. Extensive experimental results across multiple benchmarks, especially in the most limited label scenarios with the most serious class imbalance issues, demonstrate that DiffMatch performs favorably against state-of-the-art methods.

1651. Chain-of-Focus Prompting: Leveraging Sequential Visual Cues to Prompt Large Autoregressive Vision Models

链接：https://iclr.cc/virtual/2025/poster/28393 abstract：

1652. Flow: Modularized Agentic Workflow Automation

链接：https://iclr.cc/virtual/2025/poster/28132 abstract： Multi-agent frameworks powered by large language models (LLMs) have demonstrated great success in automated planning and task execution. However, the effective adjustment of agentic workflows during execution has not been well studied. An effective workflow adjustment is crucial in real-world scenarios, as the initial plan must adjust to unforeseen challenges and changing conditions in real time to ensure the efficient execution of complex tasks. In this paper, we define workflows as an activity-on-vertex (AOV) graph, which allows continuous workflow refinement by LLM agents through dynamic subtask allocation adjustment based on historical performance and previous AOVs. To further enhance framework performance, we emphasize modularity in workflow design based on evaluating parallelism and dependency complexity. With this design, our proposed multi-agent framework achieves efficient concurrent execution of subtasks, effective goal achievement, and enhanced error tolerance. Empirical results across various practical tasks demonstrate significant improvements in the efficiency of multi-agent frameworks through dynamic workflow refinement and modularization.

1653. From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks

链接：https://iclr.cc/virtual/2025/poster/29214 abstract： Biological and artificial neural networks develop internal representations that enable them to perform complex tasks. In artificial networks, the effectiveness of these models relies on their ability to build task specific representation, a process influenced by interactions among datasets, architectures, initialization strategies, and optimization algorithms. Prior studies highlight that different initializations can place networks in either a lazy regime, where representations remain static, or a rich/feature learning regime, where representations evolve dynamically. Here, we examine how initialization influences learning dynamics in deep linear neural networks, deriving exact solutions for lambda-balanced initializations-defined by the relative scale of weights across layers. These solutions capture the evolution of representations and the Neural Tangent Kernel across the spectrum from the rich to the lazy regimes. Our findings deepen the theoretical understanding of the impact of weight initialization on learning regimes, with implications for continual learning, reversal learning, and transfer learning, relevant to both neuroscience and practical applications.

1654. Stochastic Polyak Step-sizes and Momentum: Convergence Guarantees and Practical Performance

链接：https://iclr.cc/virtual/2025/poster/28388 abstract： Stochastic gradient descent with momentum, also known as Stochastic Heavy Ball method (SHB), is one of the most popular algorithms for solving large-scale stochastic optimization problems in various machine learning tasks. In practical scenarios, tuning the step-size and momentum parameters of the method is a prohibitively expensive and time-consuming process. In this work, inspired by the recent advantages of stochastic Polyak step-size in the performance of stochastic gradient descent (SGD), we propose and explore new Polyak-type variants

suitable for the update rule of the SHB method. In particular, using the Iterate Moving Average (IMA) viewpoint of SHB, we propose and analyze three novel step-size selections: MomSPSmax, MomDecSPS, and MomAdaSPS. For MomSPSmax, we provide convergence guarantees for SHB to a neighborhood of the solution for convex and smooth problems (without assuming interpolation). If interpolation is also satisfied, then using MomSPSmax, SHB converges to the true solution at a fast rate matching the deterministic HB. The other two variants, MomDecSPS and MomAdaSPS, are the first adaptive step-size for SHB that guarantee convergence to the exact minimizer - without a priori knowledge of the problem parameters and without assuming interpolation. Our convergence analysis of SHB is tight and obtains the convergence guarantees of stochastic Polyak step-size for SGD as a special case. We supplement our analysis with experiments validating our theory and demonstrating the effectiveness and robustness of our algorithms.

1655. Prototype antithesis for biological few-shot class-incremental learning

链接：https://iclr.cc/virtual/2025/poster/29109 abstract： Deep learning has become essential in the biological species recognition task. However, a significant challenge is the ability to continuously learn new or mutated species with limited annotated samples. Since species within the same family typically share similar traits, distinguishing between new and existing (old) species during incremental learning often faces the issue of species confusion. This can result in "catastrophic forgetting" of old species and poor learning of new ones. To address this issue, we propose a Prototype Antithesis (PA) method, which leverages the hierarchical structures in biological taxa to reduce confusion between new and old species. PA operates in two steps: Residual Prototype Learning (RPL) and Residual Prototype Mixing (RPM). RPL enables the model to learn unique prototypes for each species alongside residual prototypes representing shared traits within families. RPM generates synthetic samples by blending features of new species with residual prototypes of old species, encouraging the model to focus on species-unique traits and minimize species confusion. By integrating RPL and RPM, the proposed PA method mitigates "catastrophic forgetting" while improving generalization to new species. Extensive experiments on CUB200, PlantVillage, and Tree-of-Life datasets demonstrate that PA significantly reduces inter-species confusion and achieves state-of-the-art performance, highlighting its potential for deep learning in biological data analysis.

1656. PointOBB-v2: Towards Simpler, Faster, and Stronger Single Point Supervised Oriented Object Detection

链接：https://iclr.cc/virtual/2025/poster/29665 abstract： Single point supervised oriented object detection has gained attention and made initial progress within the community. Diverse from those approaches relying on one-shot samples or powerful pretrained models (e.g. SAM), PointOBB has shown promise due to its prior-free feature. In this paper, we propose PointOBB-v2, a simpler, faster, and stronger method to generate pseudo rotated boxes from points without relying on any other prior. Specifically, we first generate a Class Probability Map (CPM) by training the network with non-uniform positive and negative sampling. We show that the CPM is able to learn the approximate object regions and their contours. Then, Principal Component Analysis (PCA) is applied to accurately estimate the orientation and the boundary of objects. By further incorporating a separation mechanism, we resolve the confusion caused by the overlapping on the CPM, enabling its operation in high-density scenarios. Extensive comparisons demonstrate that our method achieves a training speed 15.58$\times$ faster and an accuracy improvement of 11.60\%/25.15\%/21.19\% on the DOTA-v1.0/v1.5/v2.0 datasets compared to the previous state-of-the-art, PointOBB. This significantly advances the cutting edge of single point supervised oriented detection in the modular track. Code and models will be released.

1657. Towards Explaining the Power of Constant-depth Graph Neural Networks for Structured Linear Programming

链接：https://iclr.cc/virtual/2025/poster/30174 abstract： Graph neural networks (GNNs) have recently emerged as powerful tools for solving complex optimization problems, often being employed to approximate solution mappings. Empirical evidence shows that even shallow GNNs (with fewer than ten layers) can achieve strong performance in predicting optimal solutions to linear programming (LP) problems. This finding is somewhat counter-intuitive, as LPs are global optimization problems, while shallow GNNs predict based on local information. Although previous theoretical results suggest that GNNs have the expressive power to solve LPs, they require deep architectures whose depth grows at least polynomially with the problem size, and thus leave the underlying principle of this empirical phenomenon still unclear. In this paper, we examine this phenomenon through the lens of distributed computing and average-case analysis. We establish that the expressive power of GNNs for LPs is closely related to well-studied distributed algorithms for LPs. Specifically, we show that any d-round distributed LP algorithm can be simulated by a d-depth GNN, and vice versa. In particular, by designing a new distributed LP algorithm and then unrolling it, we prove that constant-depth, constant-width GNNs suffice to solve sparse binary LPs effectively. Here, in contrast with previous analyses focusing on worst-case scenarios, in which we show that GNN depth must increase with problem size by leveraging an impossibility result about distributed LP algorithms, our analysis shifts the focus to the average-case performance, and shows that constant GNN depth then becomes sufficient no matter how large the problem size is. Our theory is validated by numerical results.

1658. On the Optimization Landscape of Low Rank Adaptation Methods for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28273 abstract： Training Large Language Models (LLMs) poses significant memory challenges, making low-rank adaptation methods an attractive solution. Previously, Low-Rank Adaptation (LoRA) addressed this by adding a trainable low-rank matrix to the frozen pre-trained weights in each layer, reducing the number of trainable parameters and optimizer states. GaLore, which compresses the gradient matrix instead of the weight matrix, has demonstrated superior performance to LoRA with faster convergence and reduced memory consumption. Despite their empirical success, the performance of these methods has not been fully understood or explained theoretically. In this paper, we analyze the optimization landscapes of LoRA, GaLore, and full-rank methods, revealing that GaLore benefits from fewer spurious local minima and a larger region that satisfies the \pl, a variant of Polyak-Łojasiewicz (PL) condition, leading to faster convergence. Our analysis leads to a novel method, GaRare, which further improves GaLore by using gradient random projection to reduce computational overhead. Practically, GaRare achieves strong performance in both pre-training and fine-tuning tasks, offering a more efficient approach to large-scale model adaptation.

1659. Multi-modal Learning: A Look Back and the Road Ahead

链接：https://iclr.cc/virtual/2025/poster/31355 abstract： Advancements in language models has spurred an increasing interest in multi-modal AI — models that process and understand information across multiple forms of data, such as text, images and audio. While the goal is to emulate human-like ability to handle diverse information, a key question is: do human-defined modalities align with machine perception? If not, how does this misalignment affect AI performance? In this blog, we examine these questions by reflecting on the progress made by the community in developing multi-modal benchmarks and architectures, highlighting their limitations. By reevaluating our definitions and assumptions, we propose ways to better handle multi-modal data by building models that analyze and combine modality contributions both independently and jointly with other modalities.

1660. Execution-guided within-prompt search for programming-by-example

链接：https://iclr.cc/virtual/2025/poster/29749 abstract： Large language models (LLMs) can generate code from examples without being limited to a DSL, but they lack search, as sampled programs are independent.In this paper, we use an LLM as a policy that generates lines of code and then join these lines of code to let the LLM implicitly estimate the value of each of these lines in its next iteration.We further guide the policy and value estimation by executing each line and annotating it with its results on the given examples. This allows us to search for programs within a single (expanding) prompt until a sound program is found, by letting the policy reason in both the syntactic (code) and semantic (execution) space.We evaluate within-prompt search on straight-line Python code generation using five benchmarks across different domains (strings, lists, and arbitrary Python programming problems).We show that the model uses the execution results to guide the search and that within-prompt search performs well at low token budgets.We also analyze how the model behaves as a policy and value, show that it can parallelize the search, and that it can implicitly backtrack over earlier generations.

1661. Metric-Driven Attributions for Vision Transformers

链接：https://iclr.cc/virtual/2025/poster/28207 abstract： Attribution algorithms explain computer vision models by attributing the model response to pixels within the input. Existing attribution methods generate explanations by combining transformations of internal model representations such as class activation maps, gradients, attention, or relevance scores. The effectiveness of an attribution map is measured using attribution quality metrics. This leads us to pose the following question: if attribution methods are assessed using attribution quality metrics, why are the metrics not used to generate the attributions? In response to this question, we propose a Metric-Driven Attribution for explaining Vision Transformers (ViT) called MDA. Guided by attribution quality metrics, the method creates attribution maps by performing patch order and patch magnitude optimization across all patch tokens. The first step orders the patches in terms of importance and the second step assigns the magnitude to each patch while preserving the patch order. Moreover, MDA can provide a smooth trade-off between sparse and dense attributions by modifying the optimization objective. Experimental evaluation demonstrates the proposed MDA method outperforms 7 existing ViT attribution methods by an average of 12% across 12 attribution metrics on the ImageNet dataset for the ViT-base 16×16, ViT-tiny 16×16, and ViT-base 32×32 models. Code is publicly available at https://github.com/chasewalker26/MDA-Metric-Driven-Attributions-for-ViT.

1662. Correlation and Navigation in the Vocabulary Key Representation Space of Language Models

链接：https://iclr.cc/virtual/2025/poster/29402 abstract： Language model (LM) decoding is based on the next-token prediction (NTP) probability distribution. For neural LMs (e.g., Transformer-based), NTP distribution isessentially a softmax-regularized dot product between an encoded input context(query) and fixed vocabulary representations (keys). In this paper, we study theeffect of the key distribution on the NTP distribution, with a focus on whetherthe similarity between keys will trigger spurious correlations in NTP. Throughknowledge-probing tasks, we show that in the NTP distribution, the few top-rankedtokens are typically accurate. However, the middle-ranked prediction is highly biasedtowards the tokens that are distributionally (not necessarily semantically) similar tothese top ones. For instance, if “P” is predicted as the top-1 token, “A”-“Z” will allbe ranked high in NTP, no matter whether they can lead to correct decoding results.This hurts the sampling diversity and makes the sampling of correct, long-tailresults hopeless and noisy. We attempt to alleviate this issue via a novel in-contextmethod that iteratively pushes the query representation away from explored regions.Specifically, we include the explored decoding results in the context and promptthe LM to generate something else, which encourages the LM to produce a queryrepresentation that has small dot products with explored keys. Experiments onknowledge-probing tasks show that our method leads to efficient

navigation awayfrom explored keys to correct new keys. We further extend our method to open-ended and chain-of-thought (for reasoning) generation. Experiment results showthat ICN contributes to better generation diversity and improved self-consistencyvoting performance. Finally, we discuss potential training issues caused by thefixed key space together with the challenges and possible ways to address them infuture research.

1663. LLMOPT: Learning to Define and Solve General Optimization Problems from Scratch

链接：https://iclr.cc/virtual/2025/poster/30695 abstract： Optimization problems are prevalent across various scenarios. Formulating and then solving optimization problems described by natural language often requires highly specialized human expertise, which could block the widespread application of optimization-based decision making. To automate problem formulation and solving, leveraging large language models (LLMs) has emerged as a potential way. However, this kind of approach suffers from the issue of optimization generalization. Namely, the accuracy of most current LLM-based methods and the generality of optimization problem types that they can model are still limited. In this paper, we propose a unified learning-based framework called LLMOPT to boost optimization generalization. Starting from the natural language descriptions of optimization problems and a pre-trained LLM, LLMOPT constructs the introduced five-element formulation as a universal model for learning to define diverse optimization problem types. Then, LLMOPT employs the multi-instruction tuning to enhance both problem formalization and solver code generation accuracy and generality. After that, to prevent hallucinations in LLMs, such as sacrificing solving accuracy to avoid execution errors, the model alignment and self-correction mechanism are adopted in LLMOPT. We evaluate the optimization generalization ability of LLMOPT and compared methods across six real-world datasets covering roughly 20 fields such as health, environment, energy and manufacturing, etc. Extensive experiment results show that LLMOPT is able to model various optimization problem types such as linear/nonlinear programming, mixed integer programming, and combinatorial optimization, and achieves a notable 11.08% average solving accuracy improvement compared with the state-of-the-art methods. The code is available at https://github.com/caigaojiang/LLMOPT.

1664. OPTAMI: Global Superlinear Convergence of High-order Methods

链接：https://iclr.cc/virtual/2025/poster/30494 abstract： Second-order methods for convex optimization outperform first-order methods in terms of theoretical iteration convergence, achieving rates up to $O(k^{-5})$ for highly-smooth functions. However, their practical performance and applications are limited due to their multi-level structure and implementation complexity. In this paper, we present new results on high-order optimization methods, supported by their practical performance. First, we show that the basic high-order methods, such as the Cubic Regularized Newton Method, exhibit global superlinear convergence for μ-strongly star-convex functions, a class that includes μ-strongly convex functions and some non-convex functions. Theoretical convergence results are both inspired and supported by the practical performance of these methods. Secondly, we propose a practical version of the Nesterov Accelerated Tensor method, called NATA. It significantly outperforms the classical variant and other high-order acceleration techniques in practice. The convergence of NATA is also supported by theoretical results. Finally, we introduce an open-source computational library for high-order methods, called OPTAMI. This library includes various methods, acceleration techniques, and subproblem solvers, all implemented as PyTorch optimizers, thereby facilitating the practical application of high-order methods to a wide range of optimization problems. We hope this library will simplify research and practical comparison of methods beyond first-order.

1665. SOO-Bench: Benchmarks for Evaluating the Stability of Offline Black-Box Optimization

链接：https://iclr.cc/virtual/2025/poster/29089 abstract： Black-box optimization aims to find the optima through building a model close to the black-box objective function based on function value evaluation. However, in many real-world tasks, such as the design of molecular formulas and mechanical structures, it is perilous, costly, or even infeasible to evaluate the objective function value of an actively sampled solution. In this situation, optimization can only be conducted via utilizing offline historical data, which yields offline black-box optimization. Different from the traditional goal that is to pursue the optimal solution, this paper emphasizes that the goal of offline optimization is to stably surpass the offline dataset during optimization procedure. Although benchmarks called Design-Bench already exist in this emerging field, it can hardly evaluate the stability of offline optimization and mainly provides real-world offline tasks and the corresponding offline datasets. To this end, this paper proposes benchmarks named SOO-Bench (i.e., Stable Offline Optimization Benchmarks) for offline black-box optimization algorithms, so as to systematically evaluate the stability of surpassing the offline dataset under different data distributions. Along with SOO-Bench, we also propose a stability indicator to measure the degree of stability. Specifically, SOO-Bench includes various real-world offline optimization tasks and offline datasets under different data distributions, involving the fields of satellites, materials science, structural mechanics, and automobile manufacturing. Empirically, baseline and state-of-the-art algorithms are tested and analyzed on SOO-Bench. Hopefully, SOO-Bench is expected to serve as a catalyst for the rapid developments of more novel and stable offline optimization methods. The code is available at \url{https://github.com/zhuyiyi-123/SOO-Bench}.

1666. Centrality-guided Pre-training for Graph

链接：https://iclr.cc/virtual/2025/poster/29321 abstract： Self-supervised learning (SSL) has shown great potential in learning generalizable representations for graph-structured data. However, existing SSL-based graph pre-training methods largely focus on improving graph representations by learning the structure information based on disturbing or reconstructing graphs, which

ignores an important issue: the importance of different nodes in the graph structure may vary. To fill this gap, we propose a Centrality-guided Graph Pre-training (CenPre) framework to integrate the distinct importance of nodes in graph structure into the corresponding representations of nodes based on the centrality in graph theory. In this way, the different roles played by different nodes can be effectively leveraged when learning graph structure. The proposed CenPre contains three modules for node representation pre-training and alignment. The node-level importance learning module fuses the fine-grained node importance into node representation based on degree centrality, allowing the aggregation of node representations with equal/similar importance. The graph-level importance learning module characterizes the importance between all nodes in the graph based on eigenvector centrality, enabling the exploitation of graph-level structure similarities/differences when learning node representation. Finally, a representation alignment module aligns the pre-trained node representation using the original one, essentially allowing graph representations to learn structural information without losing their original semantic information, thereby leading to better graph representations. Extensive experiments on a series of real-world datasets demonstrate that the proposed CenPre outperforms the state-of-the-art baselines in the tasks of node classification, link prediction, and graph classification.

1667. Emergent Orientation Maps —— Mechanisms, Coding Efficiency and Robustness

链接：https://iclr.cc/virtual/2025/poster/28157 abstract： Extensive experimental studies have shown that in lower mammals, neuronal orientation preference in the primary visual cortex is organized in disordered "salt-and-pepper" organizations. In contrast, higher-order mammals display a continuous variation in orientation preference, forming pinwheel-like structures. Despite these observations, the spiking mechanisms underlying the emergence of these distinct topological structures and their functional roles in visual processing remain poorly understood. To address this, we developed a self-evolving spiking neural network model with Hebbian plasticity, trained using physiological parameters characteristic of rodents, cats, and primates, including retinotopy, neuronal morphology, and connectivity patterns. Our results identify critical factors, such as the degree of input visual field overlap, neuronal connection range, and the balance between localized connectivity and long-range competition, that determine the emergence of either salt-and-pepper or pinwheel-like topologies. Furthermore, we demonstrate that pinwheel structures exhibit lower wiring costs and enhanced sparse coding capabilities compared to salt-and-pepper organizations. They also maintain greater coding robustness against noise in naturalistic visual stimuli. These findings suggest that such topological structures confer significant computational advantages in visual processing and highlight their potential application in the design of brain-inspired deep learning networks and algorithms.

1668. Unveiling the Magic of Code Reasoning through Hypothesis Decomposition and Amendment

链接：https://iclr.cc/virtual/2025/poster/28593 abstract： The reasoning abilities are one of the most enigmatic and captivating aspects of large language models (LLMs). Numerous studies are dedicated to exploring and expanding the boundaries of this reasoning capability. However, tasks that embody both reasoning and recall characteristics are often overlooked. In this paper, we introduce such a novel task, code reasoning, to provide a new perspective for the reasoning abilities of LLMs.We summarize three meta-benchmarks based on established forms of logical reasoning, and instantiate these into eight specific benchmark tasks. Our testing on these benchmarks reveals that LLMs continue to struggle with identifying satisfactory reasoning pathways. Additionally, we present a new pathway exploration pipeline inspired by human intricate problem-solving methods. This Reflective Hypothesis Decomposition and Amendment (RHDA) pipeline consists of the following iterative steps: (1) Proposing potential hypotheses based on observations and decomposing them; (2) Utilizing tools to validate hypotheses and reflection outcomes; (3) Revising hypothesis in light of observations. Our approach effectively mitigates logical chain collapses arising from forgetting or hallucination issues in multi-step reasoning, resulting in performance gains of up to $3\times$. Finally, we expanded this pipeline by applying it to simulate complex household tasks in real-world scenarios, specifically in VirtualHome, enhancing the handling of failure cases. We release our code and all of results at https://github.com/TnTWoW/code_reasoning.

1669. Do not write that jailbreak paper

链接：https://iclr.cc/virtual/2025/poster/31351 abstract： Jailbreaks are becoming a new ImageNet competition instead of helping us better understand LLM security. This blogpost surveys the jailbreak literature to extract the most important contributions and encourages the community to revisit their choices and focus on research that can uncover new security vulnerabilities.

1670. Dynamic Gaussians Mesh: Consistent Mesh Reconstruction from Dynamic Scenes

链接：https://iclr.cc/virtual/2025/poster/29972 abstract： Modern 3D engines and graphics pipelines require mesh as a memory-efficient representation, which allows efficient rendering, geometry processing, texture editing, and many other downstream operations. However, it is still highly difficult to obtain high-quality mesh in terms of detailed structure and time consistency from dynamic observations. To this end, we introduce Dynamic Gaussians Mesh (DG-Mesh), a framework to reconstruct a high-fidelity and time-consistent mesh from dynamic input. Our work leverages the recent advancement in 3D Gaussian Splatting to construct the mesh sequence with temporal consistency from dynamic observations. Building on top of this

representation, DG-Mesh recovers high-quality meshes from the Gaussian points and can track the mesh vertices over time, which enables applications such as texture editing on dynamic objects. We introduce the Gaussian-Mesh Anchoring, which encourages evenly distributed Gaussians, resulting better mesh reconstruction through mesh-guided densification and pruning on the deformed Gaussians. By applying cycle-consistent deformation between the canonical and the deformed space, we can project the anchored Gaussian back to the canonical space and optimize Gaussians across all time frames. During the evaluation on different datasets, DG-Mesh provides significantly better mesh reconstruction and rendering than baselines.

1671. It Helps to Take a Second Opinion: Teaching Smaller LLMs To Deliberate Mutually via Selective Rationale Optimisation

链接：https://iclr.cc/virtual/2025/poster/29886 abstract： Very large language models (LLMs) such as GPT-4 have shown the ability to handle complex tasks by generating and self-refining step-by-step rationales. Smaller language models (SLMs), typically with < 13B parameters, have been improved by using the data generated from very-large LMs through knowledge distillation. However, various practical constraints such as API costs, copyright, legal and ethical policies restrict using large (often opaque) models to train smaller models for commercial use. Limited success has been achieved at improving the ability of an SLM to explore the space of possible rationales and evaluate them by itself through self-deliberation. To address this, we propose COALITION, a trainable framework that facilitates interaction between two variants of the same SLM and trains them to generate and refine rationales optimized for the end-task. The variants exhibit different behaviors to produce a set of diverse candidate rationales during the generation and refinement steps. The model is then trained via Selective Rationale Optimization (SRO) to prefer generating rationale candidates that maximize the likelihood of producing the ground-truth answer. During inference, COALITION employs a controller to select the suitable variant for generating and refining the rationales. On five different datasets covering mathematical problems, commonsense reasoning, and natural language inference, COALITION outperforms several baselines by up to 5%. Our ablation studies reveal that cross-communication between the two variants performs better than using the single model to self-refine the rationales. We also demonstrate the applicability of COALITION for LMs of varying scales (4B to 14B parameters) and model families (Mistral, Llama, Qwen, Phi). We release the code for this work here.

1672. Symbolic regression via MDLformer-guided search: from minimizing prediction error to minimizing description length

链接：https://iclr.cc/virtual/2025/poster/28504 abstract： Symbolic regression, a task discovering the formula best fitting the given data, is typically based on the heuristical search. These methods usually update candidate formulas to obtain new ones with lower prediction errors iteratively. However, since formulas with similar function shapes may have completely different symbolic forms, the prediction error does not decrease monotonously as the search approaches the target formula, causing the low recovery rate of existing methods. To solve this problem, we propose a novel search objective based on the minimum description length, which reflects the distance from the target and decreases monotonically as the search approaches the correct form of the target formula. To estimate the minimum description length of any input data, we design a neural network, MDLformer, which enables robust and scalable estimation through large-scale training. With the MDLformer's output as the search objective, we implement a symbolic regression method, SR4MDL, that can effectively recover the correct mathematical form of the formula. Extensive experiments illustrate its excellent performance in recovering formulas from data. Our method successfully recovers around 50 formulas across two benchmark datasets comprising 133 problems, outperforming state-of-the-art methods by 43.92%. Experiments on 122 unseen black-box problems further demonstrate its generalization performance. We release our code at https://github.com/tsinghua-fib-lab/SR4MDL .

1673. Learning Partial Graph Matching via Optimal Partial Transport

链接：https://iclr.cc/virtual/2025/poster/27991 abstract： Partial graph matching extends traditional graph matching by allowing some nodes to remain unmatched, enabling applications in more complex scenarios. However, this flexibility introduces additional complexity, as both the subset of nodes to match and the optimal mapping must be determined. While recent studies have explored deep learning techniques for partial graph matching, a significant limitation remains: the absence of an optimization objective that fully captures the problem’s intrinsic nature while enabling efficient solutions. In this paper, we propose a novel optimization framework for partial graph matching, inspired by optimal partial transport. Our approach formulates an objective that enables partial assignments while incorporating matching biases, using weighted total variation as the divergence function to guarantee optimal partial assignments. Our method can achieve efficient, exact solutions within cubic worst case time complexity. Our contributions are threefold: (i) we introduce a novel optimization objective that balances matched and unmatched nodes; (ii) we establish a connection between partial graph matching and linear sum assignment problem, enabling efficient solutions; (iii) we propose a deep graph matching architecture with a novel partial matching loss, providing an end-to-end solution. The empirical evaluations on standard graph matching benchmarks demonstrate the efficacy of the proposed approach.

1674. ComLoRA: A Competitive Learning Approach for Enhancing LoRA

链接：https://iclr.cc/virtual/2025/poster/28642 abstract： We propose a Competitive Low-Rank Adaptation (ComLoRA) framework to address the limitations of the LoRA method, which either lacks capacity with a single rank-r LoRA or risks inefficiency and overfitting with a larger rank-Kr LoRA, where K is an integer larger than 1. The proposed ComLoRA

method initializes K distinct LoRA components, each with rank r, and allows them to compete during training. This competition drives each LoRA component to outperform the others, improving overall model performance. The best-performing LoRA is selected based on validation metrics, ensuring that the final model outperforms a single rank-r LoRA and matches the effectiveness of a larger rank-Kr LoRA, all while avoiding extra computational overhead during inference. To the best of our knowledge, this is the first work to introduce and explore competitive learning in the context of LoRA optimization. The ComLoRA's code is available at https://github.com/hqsiswiliam/comlora.

1675. Learnable Expansion of Graph Operators for Multi-Modal Feature Fusion

链接：https://iclr.cc/virtual/2025/poster/29612 abstract： In computer vision tasks, features often come from diverse representations, domains (e.g., indoor and outdoor), and modalities (e.g., text, images, and videos). Effectively fusing these features is essential for robust performance, especially with the availability of powerful pre-trained models like vision-language models. However, common fusion methods, such as concatenation, element-wise operations, and non-linear techniques, often fail to capture structural relationships, deep feature interactions, and suffer from inefficiency or misalignment of features across domains or modalities. In this paper, we shift from high-dimensional feature space to a lower-dimensional, interpretable graph space by constructing relationship graphs that encode feature relationships at different levels, e.g., clip, frame, patch, token, etc. To capture deeper interactions, we expand graphs through iterative graph relationship updates and introduce a learnable graph fusion operator to integrate these expanded relationships for more effective fusion. Our approach is relationship-centric, operates in a homogeneous space, and is mathematically principled, resembling element-wise relationship score aggregation via multilinear polynomials. We demonstrate the effectiveness of our graph-based fusion method on video anomaly detection, showing strong performance across multi-representational, multi-modal, and multi-domain feature fusion tasks.

1676. MTSAM: Multi-Task Fine-Tuning for Segment Anything Model

链接：https://iclr.cc/virtual/2025/poster/30892 abstract： The Segment Anything Model (SAM), with its remarkable zero-shot capability, has the potential to be a foundation model for multi-task learning. However, adopting SAM to multi-task learning faces two challenges: (a) SAM has difficulty generating task-specific outputs with different channel numbers, and (b) how to fine-tune SAM to adapt multiple downstream tasks simultaneously remains unexplored. To address these two challenges, in this paper, we propose the Multi-Task SAM (MTSAM) framework, which enables SAM to work as a foundation model for multi-task learning. MTSAM modifies SAM's architecture by removing the prompt encoder and implementing task-specific no-mask embeddings and mask decoders, enabling the generation of task-specific outputs. Furthermore, we introduce Tensorized low-Rank Adaptation (ToRA) to perform multi-task fine-tuning on SAM. Specifically, ToRA injects an update parameter tensor into each layer of the encoder in SAM and leverages a low-rank tensor decomposition method to incorporate both task-shared and task-specific information.Extensive experiments conducted on benchmark datasets substantiate the efficacy of MTSAM in enhancing the performance of multi-task learning. Our code is available at https://github.com/XuehaoWangFi/MTSAM.

1677. HeadMap: Locating and Enhancing Knowledge Circuits in LLMs

链接：https://iclr.cc/virtual/2025/poster/28637 abstract： Large language models (LLMs), through pretraining on extensive corpora, encompass rich semantic knowledge and exhibit the potential for efficient adaptation to diverse downstream tasks. However, the intrinsic mechanisms underlying LLMs remain unexplored, limiting the efficacy of applying these models to downstream tasks. In this paper, we explore the intrinsic mechanisms of LLMs from the perspective of knowledge circuits. Specifically, considering layer dependencies, we propose a layer-conditioned locating algorithm to identify a series of attention heads, which is a knowledge circuit of some tasks. Experiments demonstrate that simply masking a small portion of attention heads in the knowledge circuit can significantly reduce the model's ability to make correct predictions. This suggests that the knowledge flow within the knowledge circuit plays a critical role when the model makes a correct prediction. Inspired by this observation, we propose a novel parameter-efficient fine-tuning method called HeadMap, which maps the activations of these critical heads in the located knowledge circuit to the residual stream by two linear layers, thus enhancing knowledge flow from the knowledge circuit in the residual stream. Extensive experiments conducted on diverse datasets demonstrate the efficiency and efficacy of the proposed method. Our code is available at https://github.com/XuehaoWangFi/HeadMap.

1678. Learning to Explore and Exploit with GNNs for Unsupervised Combinatorial Optimization

链接：https://iclr.cc/virtual/2025/poster/27904 abstract： Combinatorial optimization (CO) problems are pervasiveacross various domains, but their NP-hard nature often necessitates problem-specificheuristic algorithms. Recent advancements in deep learning have led to the development of learning-based heuristics, yet these approaches often struggle with limited search capabilities.We introduce Explore-and-Exploit GNN (X^2GNN, pronounced x-squared GNN), a novel unsupervised neural framework that combines exploration and exploitation for combinatorial search optimization:i) Exploration - X^2GNN generates multiple solutions simultaneously, promoting diversity in the search space; (ii) Exploitation - X^2GNN employs neural stochastic iterative refinement to exploit partial existing solutions, guiding the search toward promising regions and helping escape local optima.By balancing exploration and exploitation, X^2GNN achieves superior performance and generalization on several graph CO problems including Max Cut, Max Independent Set, and Max Clique. Notably, for large Max Clique problems, X^2GNN consistently generates solutions within 1.2\% of optimality, while other state-of-the-art learning-

based approaches struggle to reach within 22\% of optimal. Moreover, X^2GNN consistently generates better solutions than Gurobi on large graphs for all three problems under reasonable time budgets. Furthermore, X^2GNN exhibits exceptional generalization capabilities. For the Maximum Independent Set problem, X^2GNN outperforms state-of-the-art methods even when trained on smaller or out-of-distribution graphs compared to the test set. Our framework offers a more effective and flexible approach to neural combinatorial optimization, addressing a key challenge in the field and providing a promising direction for future research in learning-based heuristics for combinatorial optimization.

1679. Trajectory-Class-Aware Multi-Agent Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/27954 abstract： In the context of multi-agent reinforcement learning, generalization is a challenge to solve various tasks that may require different joint policies or coordination without relying on policies specialized for each task. We refer to this type of problem as a multi-task, and we train agents to be versatile in this multi-task setting through a single training process. To address this challenge, we introduce TRajectory-class-Aware Multi-Agent reinforcement learning (TRAMA). In TRAMA, agents recognize a task type by identifying the class of trajectories they are experiencing through partial observations, and the agents use this trajectory awareness or prediction as additional information for action policy. To this end, we introduce three primary objectives in TRAMA: (a) constructing a quantized latent space to generate trajectory embeddings that reflect key similarities among them; (b) conducting trajectory clustering using these trajectory embeddings; and (c) building a trajectory-class-aware policy. Specifically for (c), we introduce a trajectory-class predictor that performs agent-wise predictions on the trajectory class; and we design a trajectory-class representation model for each trajectory class. Each agent takes actions based on this trajectory-class representation along with its partial observation for task-aware execution. The proposed method is evaluated on various tasks, including multi-task problems built upon StarCraft II. Empirical results show further performance improvements over state-of-the-art baselines.

1680. Generalized Video Moment Retrieval

链接：https://iclr.cc/virtual/2025/poster/28243 abstract： In this paper, we introduce the Generalized Video Moment Retrieval (GVMR) framework, which extends traditional Video Moment Retrieval (VMR) to handle a wider range of query types. Unlike conventional VMR systems, which are often limited to simple, single-target queries, GVMR accommodates both non-target and multi-target queries. To support this expanded task, we present the NExT-VMR dataset, derived from the YFCC100M collection, featuring diverse query scenarios to enable more robust model evaluation.Additionally, we propose BCANet, a transformer-based model incorporating the novel Boundary-aware Cross Attention (BCA) module. The BCA module enhances boundary detection and uses cross-attention to achieve a comprehensive understanding of video content in relation to queries. BCANet accurately predicts temporal video segments based on natural language descriptions, outperforming traditional models in both accuracy and adaptability. Our results demonstrate the potential of the GVMR framework, the NExT-VMR dataset, and BCANet to advance VMR systems, setting a new standard for future multimedia information retrieval research.

1681. Modeling Unseen Environments with Language-guided Composable Causal Components in Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29309 abstract： Generalization in reinforcement learning (RL) remains a significant challenge, especially when agents encounter novel environments with unseen dynamics. Drawing inspiration from human compositional reasoning—where known components are reconfigured to handle new situations—we introduce World Modeling with Compositional Causal Components (WM3C). This novel framework enhances RL generalization by learning and leveraging compositional causal components. Unlike previous approaches focusing on invariant representation learning or meta-learning, WM3C identifies and utilizes causal dynamics among composable elements, facilitating robust adaptation to new tasks. Our approach integrates language as a compositional modality to decompose the latent space into meaningful components and provides theoretical guarantees for their unique identification under mild assumptions. Our practical implementation uses a masked autoencoder with mutual information constraints and adaptive sparsity regularization to capture high-level semantic information and effectively disentangle transition dynamics. Experiments on numerical simulations and real-world robotic manipulation tasks demonstrate that WM3C significantly outperforms existing methods in identifying latent processes, improving policy learning, and generalizing to unseen tasks.

1682. HiRA: Parameter-Efficient Hadamard High-Rank Adaptation for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29500 abstract： We propose Hadamard High-Rank Adaptation (HiRA), a parameter-efficient fine-tuning (PEFT) method that enhances the adaptability of Large Language Models (LLMs). While Low-rank Adaptation (LoRA) is widely used to reduce resource demands, its low-rank updates may limit its expressiveness for new tasks. HiRA addresses this by using a Hadamard product to retain high-rank update parameters, improving the model capacity. Empirically, HiRA outperforms LoRA and its variants on several tasks, with extensive ablation studies validating its effectiveness. Our code is available at https://github.com/hqsiswiliam/hira.

1683. Sharpness-Aware Black-Box Optimization

链接：https://iclr.cc/virtual/2025/poster/28779 abstract： Black-box optimization algorithms have been widely used in various machine learning problems, including reinforcement learning and prompt fine-tuning. However, directly optimizing the training loss value, as commonly done in existing black-box optimization methods, could lead to suboptimal model quality and generalization performance. To address those problems in black-box optimization, we propose a novel Sharpness-Aware Black-box Optimization (SABO) algorithm, which applies a sharpness-aware minimization strategy to improve the model generalization. Specifically, the proposed SABO method first reparameterizes the objective function by its expectation over a Gaussian distribution. Then it iteratively updates the parameterized distribution by approximated stochastic gradients of the maximum objective value within a small neighborhood around the current solution in the Gaussian distribution space. Theoretically, we prove the convergence rate and generalization bound of the proposed SABO algorithm. Empirically, extensive experiments on the black-box prompt fine-tuning tasks demonstrate the effectiveness of the proposed SABO method in improving model generalization performance.

1684. Better Instruction-Following Through Minimum Bayes Risk

链接：https://iclr.cc/virtual/2025/poster/30792 abstract： General-purpose LLM judges capable of human-level evaluation provide not only a scalable and accurate way of evaluating instruction-following LLMs but also new avenues for supervising and improving their performance. One promising way of leveraging LLM judges for supervision is through Minimum Bayes Risk (MBR) decoding, which uses a reference-based evaluator to select a high-quality output from amongst a set of candidate outputs. In the first part of this work, we explore using MBR decoding as a method for improving the test-time performance of instruction-following LLMs. We find that MBR decoding with reference-based LLM judges substantially improves over greedy decoding, best-of-N decoding with reference-free judges and MBR decoding with lexical and embedding-based metrics on AlpacaEval and MT-Bench. These gains are consistent across LLMs with up to 70B parameters, demonstrating that smaller LLM judges can be used to supervise much larger LLMs. Then, seeking to retain the improvements from MBR decoding while mitigating additional test-time costs, we explore iterative self-training on MBR-decoded outputs. We find that self-training using Direct Preference Optimisation leads to significant performance gains, such that the self-trained models with greedy decoding generally match and sometimes exceed the performance of their base models with MBR decoding.

1685. AFlow: Automating Agentic Workflow Generation

链接：https://iclr.cc/virtual/2025/poster/27691 abstract： Large language models (LLMs) have demonstrated remarkable potential in solving complex tasks across diverse domains, typically by employing agentic workflows that follow detailed instructions and operational sequences. However, constructing these workflows requires significant human effort, limiting scalability and generalizability. Recent research has sought to automate the generation and optimization of these workflows, but existing methods still rely on initial manual setup and fall short of achieving fully automated and effective workflow generation. To address this challenge, we reformulate workflow optimization as a search problem over code-represented workflows, where LLM-invoking nodes are connected by edges. We introduce AFLOW, an automated framework that efficiently explores this space using Monte Carlo Tree Search, iteratively refining workflows through code modification, tree-structured experience, and execution feedback. Empirical evaluations across six benchmark datasets demonstrate AFLOW's efficacy, yielding a 5.7% average improvement over state-of-the-art baselines. Furthermore, AFLOW enables smaller models to outperform GPT-4o on specific tasks at 4.55% of its inference cost in dollars. The code is available at https://github.com/geekan/MetaGPT.

1686. Long-Short Decision Transformer: Bridging Global and Local Dependencies for Generalized Decision-Making

链接：https://iclr.cc/virtual/2025/poster/29888 abstract： Decision Transformers (DTs) effectively capture long-range dependencies using self-attention but struggle with fine-grained local relationships, especially the Markovian properties in many offline-RL datasets. Conversely, Decision Convformer (DC) utilizes convolutional filters for capturing local patterns but shows limitations in tasks demanding long-term dependencies, such as Maze2d. To address these limitations and leverage both strengths, we propose the Long-Short Decision Transformer (LSDT), a general-purpose architecture to effectively capture global and local dependencies across two specialized parallel branches (self-attention and convolution). We explore how these branches complement each other by modeling various ranged dependencies across different environments, and compare it against other baselines. Experimental results demonstrate our LSDT achieves state-of-the-art performance and notable gains over the standard DT in D4RL offline RL benchmark. Leveraging the parallel architecture, LSDT performs consistently on diverse datasets, including Markovian and non-Markovian. We also demonstrate the flexibility of LSDT's architecture, where its specialized branches can be replaced or integrated into models like DC to improve their performance in capturing diverse dependencies. Finally, we also highlight the role of goal states in improving decision-making for goal-reaching tasks like Antmaze.

1687. Training-free LLM-generated Text Detection by Mining Token Probability Sequences

链接：https://iclr.cc/virtual/2025/poster/27889 abstract： Large language models (LLMs) have demonstrated remarkable capabilities in generating high-quality texts across diverse domains. However, the potential misuse of LLMs has raised significant concerns, underscoring the urgent need for reliable detection of LLM-generated texts. Conventional training-based detectors often struggle with generalization, particularly in cross-domain and cross-model scenarios. In contrast, training-free

methods, which focus on inherent discrepancies through carefully designed statistical features, offer improved generalization and interpretability. Despite this, existing training-free detection methods typically rely on global text sequence statistics, neglecting the modeling of local discriminative features, thereby limiting their detection efficacy. In this work, we introduce a novel training-free detector, termed \textbf{Lastde}\footnote{The code and data are released at \url{https://github.com/TrustMedia-zju/Lastde_Detector}.} that synergizes local and global statistics for enhanced detection. For the first time, we introduce time series analysis to LLM-generated text detection, capturing the temporal dynamics of token probability sequences. By integrating these local statistics with global ones, our detector reveals significant disparities between human and LLM-generated texts. We also propose an efficient alternative, \textbf{Lastde++} to enable real-time detection. Extensive experiments on six datasets involving cross-domain, cross-model, and cross-lingual detection scenarios, under both white-box and black-box settings, demonstrated that our method consistently achieves state-of-the-art performance. Furthermore, our approach exhibits greater robustness against paraphrasing attacks compared to existing baseline methods.

1688. Leveraging Driver Field-of-View for Multimodal Ego-Trajectory Prediction

链接：https://iclr.cc/virtual/2025/poster/30006 abstract： Understanding drivers' decision-making is crucial for road safety. Although predicting the ego-vehicle's path is valuable for driver-assistance systems, existing methods mainly focus on external factors like other vehicles' motions, often neglecting the driver's attention and intent. To address this gap, we infer the ego-trajectory by integrating the driver's gaze and the surrounding scene. We introduce RouteFormer, a novel multimodal ego-trajectory prediction network combining GPS data, environmental context, and the driver's field-of-view—comprising first-person video and gaze fixations. We also present the Path Complexity Index (PCI), a new metric for trajectory complexity that enables a more nuanced evaluation of challenging scenarios. To tackle data scarcity and enhance diversity, we introduce GEM, a comprehensive dataset of urban driving scenarios enriched with synchronized driver field-of-view and gaze data. Extensive evaluations on GEM and DR(eye)VE demonstrate that RouteFormer significantly outperforms state-of-the-art methods, achieving notable improvements in prediction accuracy across diverse conditions. Ablation studies reveal that incorporating driver field-of-view data yields significantly better average displacement error, especially in challenging scenarios with high PCI scores, underscoring the importance of modeling driver attention. All data and code are available at meakbiyik.github.io/routeformer.

1689. Collab: Controlled Decoding using Mixture of Agents for LLM Alignment

链接：https://iclr.cc/virtual/2025/poster/30799 abstract： Alignment of Large Language models (LLMs) is crucial for safe and trustworthy deployment in applications. Reinforcement learning from human feedback (RLHF) has emerged as an effective technique to align LLMs to human preferences, and broader utilities, but it requires updating billions of model parameters which is computationally expensive. Controlled Decoding, by contrast, provides a mechanism for aligning a model at inference time without retraining. However, single-agent decoding approaches often struggle to adapt to diverse tasks due to the complexity and variability inherent in these tasks. To strengthen the test-time performance w.r.t the target task, we propose a mixture of agents-based decoding strategies leveraging the existing off-the-shelf aligned LLM policies. Treating each prior policy as an agent in the spirit of mixture of agent collaboration, we develop a decoding method that allows for inference-time alignment through a token-level selection strategy among multiple agents. For each token, the most suitable LLM is dynamically chosen from a pool of models based on a long-term utility metric. This policy-switching mechanism ensures optimal model selection at each step, enabling efficient collaboration and alignment among LLMs during decoding. Theoretical analysis of our proposed algorithm establishes optimal performance with respect to the target task represented via a target reward, for the given off-the-shelf models. We conduct comprehensive empirical evaluations with open-source aligned models on diverse tasks and preferences, which demonstrates the merits of this approach over single-agent decoding baselines. Notably, COLLAB surpasses the current SoTA decoding strategy, achieving an improvement of {up to 1.56x} in average reward and 71.89% in GPT-4 based win-tie rate.

1690. VL-Cache: Sparsity and Modality-Aware KV Cache Compression for Vision-Language Model Inference Acceleration

链接：https://iclr.cc/virtual/2025/poster/30231 abstract： Vision-Language Models (VLMs) have demonstrated impressive performance across a versatile set of tasks. A key challenge in accelerating VLMs is storing and accessing the large Key-Value (KV) cache that encodes long visual contexts, such as images or videos. While existing KV cache compression methods are effective for Large Language Models (LLMs), directly migrating them to VLMs yields suboptimal accuracy and speedup. To bridge the gap, we propose VL-Cache, a novel KV cache compression recipe tailored for accelerating VLM inference. In this paper, we first investigate the unique sparsity pattern of VLM attention by distinguishing visual and text tokens in prefill and decoding phases. Based on these observations, we introduce a layer-adaptive sparsity-aware cache budget allocation method that effectively distributes the limited cache budget across different layers, further reducing KV cache size without compromising accuracy. Additionally, we develop a modality-aware token scoring policy to better evaluate the token importance. Empirical results on multiple benchmark datasets demonstrate that retaining only 10% of KV cache achieves accuracy comparable to that with full cache. In a speed benchmark, our method accelerates end-to-end latency of generating 100 tokens by up to 2.33x and speeds up decoding by up to 7.08x, while reducing the memory footprint of KV cache in GPU by 90%.

1691. Joint Gradient Balancing for Data Ordering in Finite-Sum Multi-Objective Optimization

链接：https://iclr.cc/virtual/2025/poster/28185 abstract： In finite-sum optimization problems, the sample orders for parameter updates can significantly influence the convergence rate of optimization algorithms. While numerous sample ordering techniques have been proposed in the context of single-objective optimization, the problem of sample ordering in finite-sum multi-objective optimization has not been thoroughly explored. To address this gap, we propose a sample ordering method called JoGBa, which finds the sample orders for multiple objectives by jointly performing online vector balancing on the gradients of all objectives. Our theoretical analysis demonstrates that this approach outperforms the standard baseline of random ordering and accelerates the convergence rate for the MGDA algorithm. Empirical evaluation across various datasets with different multi-objective optimization algorithms further demonstrates that JoGBa can achieve faster convergence and superior final performance than other data ordering strategies.

1692. Advancing Out-of-Distribution Detection via Local Neuroplasticity

链接：https://iclr.cc/virtual/2025/poster/31215 abstract： In the domain of machine learning, the assumption that training and test data share the same distribution is often violated in real-world scenarios, requiring effective out-of-distribution (OOD) detection. This paper presents a novel OOD detection method that leverages the unique local neuroplasticity property of Kolmogorov-Arnold Networks (KANs). Unlike traditional multilayer perceptrons, KANs exhibit local plasticity, allowing them to preserve learned information while adapting to new tasks. Our method compares the activation patterns of a trained KAN against its untrained counterpart to detect OOD samples. We validate our approach on benchmarks from image and medical domains, demonstrating superior performance and robustness compared to state-of-the-art techniques. These results underscore the potential of KANs in enhancing the reliability of machine learning systems in diverse environments.

1693. Hierarchically Encapsulated Representation for Protocol Design in Self-Driving Labs

链接：https://iclr.cc/virtual/2025/poster/30666 abstract： Self-driving laboratories have begun to replace human experimenters in performing single experimental skills or predetermined experimental protocols. However, as the pace of idea iteration in scientific research has been intensified by Artificial Intelligence, the demand for rapid design of new protocols for new discoveries become evident. Efforts to automate protocol design have been initiated, but the capabilities of knowledge-based machine designers, such as Large Language Models, have not been fully elicited, probably for the absence of a systematic representation of experimental knowledge, as opposed to isolated, flatten pieces of information. To tackle this issue, we propose a multi-faceted, multi-scale representation, where instance actions, generalized operations, and product flow models are hierarchically encapsulated using Domain-Specific Languages. We further develop a data-driven algorithm based on non-parametric modeling that autonomously customizes these representations for specific domains. The proposed representation is equipped with various machine designers to manage protocol design tasks, including planning, modification, and adjustment. The results demonstrate that the proposed method could effectively complement Large Language Models in the protocol design process, serving as an auxiliary module in the realm of machine-assisted scientific exploration.

1694. RTop-K: Ultra-Fast Row-Wise Top-K Selection for Neural Network Acceleration on GPUs

链接：https://iclr.cc/virtual/2025/poster/29763 abstract： Abstract Top-k selection algorithms are fundamental in a wide range of applications, including high-performance computing, information retrieval, big data processing, and neural network model training. In this paper, we present RTop-K, a highly efficient parallel row-wise top-k selection algorithm specifically designed for GPUs. RTop-K leverages a binary search-based approach to optimize row-wise top-k selection, providing a scalable and accelerated solution.We conduct a detailed analysis of early stopping in our algorithm, showing that it effectively maintains the testing accuracy of neural network models while substantially improving performance. Our GPU implementation of RTop-K demonstrates superior performance over state-of-the-art row-wise top-k GPU implementations, achieving an average speed-up of up to 11.49× with early stopping and 7.29× without early stopping. Moreover, RTop-K accelerates the overall training workflow of MaxK-GNNs, delivering speed-ups ranging from 11.97% to 33.29% across different models and datasets.

1695. The Journey Matters: Average Parameter Count over Pre-training Unifies Sparse and Dense Scaling Laws

链接：https://iclr.cc/virtual/2025/poster/27967 abstract： Pruning eliminates unnecessary parameters in neural networks; it offers a promising solution to the growing computational demands of large language models (LLMs). While many focus on post-training pruning, sparse pre-training--which combines pruning and pre-training into a single phase--provides a simpler alternative. In this work, we present the first systematic exploration of optimal sparse pre-training configurations for LLMs through an examination of 80 unique pruning schedules across different sparsity levels and training durations.We find that initiating pruning at 25\% of total training compute and concluding at 75\% achieves near-optimal final evaluation loss.These findings provide valuable insights for efficient and effective sparse pre-training of LLMs.Furthermore, we propose a new scaling law that

modifies the Chinchilla scaling law to use the average parameter count over pre-training.Through empirical and theoretical validation, we demonstrate that this modified scaling law accurately models evaluation loss for both sparsely and densely pre-trained LLMs, unifying scaling laws across pre-training paradigms.Our findings indicate that while sparse pre-training achieves the same final model quality as dense pre-training for equivalent compute budgets, it provides substantial benefits through reduced model size, enabling significant potential computational savings during inference.

1696. Manifold Constraint Reduces Exposure Bias in Accelerated Diffusion Sampling

链接：https://iclr.cc/virtual/2025/poster/30921 abstract： Diffusion models have demonstrated significant potential for generating high-quality images, audio, and videos. However, their iterative inference process entails substantial computational costs, limiting practical applications. Recently, researchers have introduced accelerated sampling methods that enable diffusion models to generate samples with far fewer timesteps than those used during training. Nonetheless, as the number of sampling steps decreases, the prediction errors significantly degrade the quality of generated outputs. Additionally, the exposure bias in diffusion models further amplifies these errors. To address these challenges, we leverage a manifold hypothesis to explore the exposure bias problem in depth. Based on this geometric perspective, we propose a manifold constraint that effectively reduces exposure bias during accelerated sampling of diffusion models. Notably, our method involves no additional training and requires only minimal hyperparameter tuning. Extensive experiments demonstrate the effectiveness of our approach, achieving a FID score of 15.60 with 10-step SDXL on MS-COCO, surpassing the baseline by a reduction of 2.57 in FID.

1697. Scaling Wearable Foundation Models

链接：https://iclr.cc/virtual/2025/poster/27713 abstract： Wearable sensors have become ubiquitous thanks to a variety of health tracking features. The resulting continuous and longitudinal measurements from everyday life generate large volumes of data. However, making sense of these observations for scientific and actionable insights is non-trivial. Inspired by the empirical success of generative modeling, where large neural networks learn powerful representations from vast amounts of text, image, video, or audio data, we investigate the scaling properties of wearable sensor foundation models across compute, data, and model size. Using a dataset of up to 40 million hours of in-situ heart rate, heart rate variability, accelerometer, electrodermal activity, skin temperature, and altimeter per-minute data from over 165,000 people, we create LSM, a multimodal foundation model built on the largest wearable-signals dataset with the most extensive range of sensor modalities to date. Our results establish the scaling laws of LSM for tasks such as imputation, interpolation and extrapolation across both time and sensor modalities. Moreover, we highlight how LSM enables sample-efficient downstream learning for tasks including exercise and activity recognition.

1698. Topological Zigzag Spaghetti for Diffusion-based Generation and Prediction on Graphs

链接：https://iclr.cc/virtual/2025/poster/28459 abstract： Diffusion models have recently emerged as a new powerful machinery for generative artificial intelligence on graphs, with applications ranging from drug design to knowledge discovery. However, despite their high potential, most, if not all, existing graph diffusion models are limited in their ability to holistically describe the intrinsic higher-order topological graph properties, which obstructs model generalizability and adoption for downstream tasks. We address this fundamental challenge and extract the latent salient topological graph descriptors at different resolutions by leveraging zigzag persistence. We develop a new computationally efficient topological summary,zigzag spaghetti (ZS), which delivers the most inherent topological properties simultaneously over a sequence of graphs at multiple resolutions. We derive theoretical stability guarantees of ZS and present the first attempt to integratedynamic topological information into graph diffusion models. Our extensive experiments on graph classification and prediction tasks suggest that ZS has a high promise not only to enhance performance of graph diffusion models, with gains up 10\%, but also to substantially booster model robustness.

1699. Divergence of Neural Tangent Kernel in Classification Problems

链接：https://iclr.cc/virtual/2025/poster/29426 abstract： This paper primarily investigates the convergence of the Neural Tangent Kernel (NTK) in classification problems. This study firstly show the strictly positive definiteness of NTK of multi-layer fully connected neural networks and residual neural networks. Then, through a contradiction argument, it indicates that, during training with the cross-entropy loss function, the neural network parameters diverge due to the strictly positive definiteness of the NTK. Consequently, the empirical NTK does not consistently converge but instead diverges as time approaches infinity. This finding implies that NTK theory is not applicable in this context, highlighting significant theoretical implications for the study of neural networks in classification problems. These results can also be easily generalized to other network structures, provided that the NTK is strictly positive definite.

1700. SplatFormer: Point Transformer for Robust 3D Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/30697 abstract： 3D Gaussian Splatting (3DGS) has recently transformed photorealistic reconstruction, achieving high visual fidelity and real-time performance. However, rendering quality significantly deteriorates when test views deviate from the camera angles used during training, posing a major challenge for applications in immersive

free-viewpoint rendering and navigation. In this work, we conduct a comprehensive evaluation of 3DGS and related novel view synthesis methods under out-of-distribution (OOD) test camera scenarios. By creating diverse test cases with synthetic and real-world datasets, we demonstrate that most existing methods, including those incorporating various regularization techniques and data-driven priors, struggle to generalize effectively to OOD views. To address this limitation, we introduce SplatFormer, the first point transformer model specifically designed to operate on Gaussian splats. SplatFormer takes as input an initial 3DGS set optimized under limited training views and refines it in a single forward pass, effectively removing potential artifacts in OOD test views. To our knowledge, this is the first successful application of point transformers directly on 3DGS sets, surpassing the limitations of previous multi-scene training methods, which could handle only a restricted number of input views during inference. Our model significantly improves rendering quality under extreme novel views, achieving state-of-the-art performance in these challenging scenarios and outperforming various 3DGS regularization techniques, multi-scene models tailored for sparse view synthesis, and diffusion-based frameworks. The project url is https://sergeyprokudin.github.io/splatformer.

1701. The Optimization Landscape of SGD Across the Feature Learning Strength

链接：https://iclr.cc/virtual/2025/poster/28707 abstract： We consider neural networks (NNs) where the final layer is down-scaled by a fixed hyperparameter γ. Recent work has identified γ as controlling the strength of feature learning.As γ increases, network evolution changes from "lazy" kernel dynamics to "rich" feature-learning dynamics, with a host of associated benefits including improved performance on common tasks.In this work, we conduct a thorough empirical investigation of the effect of scaling γ across a variety of models and datasets in the online training setting.We first examine the interaction of γ with the learning rate η, identifying several scaling regimes in the γ-η plane which we explain theoretically using a simple model.We find that the optimal learning rate η^* *scales non-trivially with γ. In particular, $\eta^* \propto \gamma^2$ when $\gamma \ll 1$ and $\eta^* \propto \gamma^{2/L}$ when $\gamma \gg 1$ for a feed-forward network of depth L.Using this optimal learning rate scaling, we proceed with an empirical study of the under-explored ``ultra-rich" $\gamma \gg 1$ regime.We find that networks in this regime display characteristic loss curves, starting with a long plateau followed by a drop-off, sometimes followed by one or more additional staircase steps.We find networks of different large γ values optimize along similar trajectories up to a reparameterization of time.We further find that optimal online performance is often found at large γ and could be missed if this hyperparameter is not tuned.Our findings indicate that analytical study of the large-γ limit may yield useful insights into the dynamics of representation learning in performant models.

1702. Depth Any Video with Scalable Synthetic Data

链接：https://iclr.cc/virtual/2025/poster/32068 abstract： Video depth estimation has long been hindered by the scarcity of consistent and scalable ground truth data, leading to inconsistent and unreliable results. In this paper, we introduce Depth Any Video, a model that tackles the challenge through two key innovations. First, we develop a scalable synthetic data pipeline, capturing real-time video depth data from diverse virtual environments, yielding 40,000 video clips of 5-second duration, each with precise depth annotations. Second, we leverage the powerful priors of generative video diffusion models to handle real-world videos effectively, integrating advanced techniques such as rotary position encoding and flow matching to further enhance flexibility and efficiency. Unlike previous models, which are limited to fixed-length video sequences, our approach introduces a novel mixed-duration training strategy that handles videos of varying lengths and performs robustly across different frame rates—even on single frames. At inference, we propose a depth interpolation method that enables our model to infer high-resolution video depth across sequences of up to 150 frames. Our model outperforms all previous generative depth models in terms of spatial accuracy and temporal consistency. The code and model weights are open-sourced.

1703. Learning Multi-Index Models with Neural Networks via Mean-Field Langevin Dynamics

链接：https://iclr.cc/virtual/2025/poster/29373 abstract： We study the problem of learning multi-index models in high-dimensions using a two-layer neural network trained with the mean-field Langevin algorithm. Under mild distributional assumptions on the data, we characterize the effective dimension d_{eff} that controls both sample and computational complexity by utilizing the adaptivity of neural networks to latent low-dimensional structures. When the data exhibit such a structure, d_{eff} can be significantly smaller than the ambient dimension. We prove that the sample complexity grows almost linearly with d_{eff}, bypassing the limitations of the information and generative exponents that appeared in recent analyses of gradient-based feature learning. On the other hand, the computational complexity may inevitably grow exponentially with d_{eff} in the worst-case scenario. Motivated by improving computational complexity, we take the first steps towards polynomial time convergence of the mean-field Langevin algorithm by investigating a setting where the weights are constrained to be on a compact manifold with positive Ricci curvature, such as the hypersphere. There, we study assumptions under which polynomial time convergence is achievable, whereas similar assumptions in the Euclidean setting lead to exponential time complexity.

1704. DOPL: Direct Online Preference Learning for Restless Bandits with Preference Feedback

链接：https://iclr.cc/virtual/2025/poster/31121 abstract： Restless multi-armed bandits (RMAB) has been widely used to model constrained sequential decision making problems, where the state of each restless arm evolves according to a Markov chain and each state transition generates a scalar reward. However, the success of RMAB crucially relies on the availability and quality of reward signals. Unfortunately, specifying an exact reward function in practice can be challenging and even infeasible. In this paper, we introduce Pref-RMAB, a new RMAB model in the presence of preference signals, where the decision maker only observes pairwise preference feedback rather than scalar reward from the activated arms at each decision epoch. Preference feedback, however, arguably contains less information than the scalar reward, which makes Pref-RMAB seemingly more difficult. To address this challenge, we present a direct online preference learning (DOPL) algorithm for Pref-RMAB to efficiently explore the unknown environments, adaptively collect preference data in an online manner, and directly leverage the preference feedback for decision-makings. We prove that DOPL yields a sublinear regret. To our best knowledge, this is the first algorithm to ensure $\tilde{\mathcal{O}}(\sqrt{T\ln T})$ regret for RMAB with preference feedback. Experimental results further demonstrate the effectiveness of DOPL.

1705. SynQ: Accurate Zero-shot Quantization by Synthesis-aware Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/31109 abstract： How can we accurately quantize a pre-trained model without any data? Quantization algorithms are widely used for deploying neural networks on resource-constrained edge devices.Zero-shot Quantization (ZSQ) addresses the crucial and practical scenario where training data are inaccessible for privacy or security reasons.However, three significant challenges hinder the performance of existing ZSQ methods: 1) noise in the synthetic dataset, 2) predictions based on off-target patterns, and the 3) misguidance by erroneous hard labels.In this paper, we propose SynQ (Synthesis-aware Fine-tuning for Zero-shot Quantization),a carefully designed ZSQ framework to overcome the limitations of existing methods.SynQ minimizes the noise from the generated samples by exploiting a low-pass filter.Then, SynQ trains the quantized model to improve accuracy by aligning its class activation map with the pre-trained model.Furthermore, SynQ mitigates misguidance from the pre-trained model's error by leveraging only soft labels for difficult samples.Extensive experiments show that SynQ provides the state-of-the-art accuracy, over existing ZSQ methods.

1706. Sensitivity-Aware Amortized Bayesian Inference

链接：https://iclr.cc/virtual/2025/poster/31469 abstract： Sensitivity analyses reveal the influence of various modeling choices on the outcomes of statistical analyses. While theoretically appealing, they are overwhelmingly inefficient for complex Bayesian models. In this work, we propose sensitivity-aware amortized Bayesian inference (SA-ABI), a multifaceted approach to efficiently integrate sensitivity analyses into simulation-based inference with neural networks. First, we utilize weight sharing to encode the structural similarities between alternative likelihood and prior specifications in the training process with minimal computational overhead. Second, we leverage the rapid inference of neural networks to assess sensitivity to data perturbations and preprocessing steps. In contrast to most other Bayesian approaches, both steps circumvent the costly bottleneck of refitting the model for each choice of likelihood, prior, or data set. Finally, we propose to use deep ensembles to detect sensitivity arising from unreliable approximation (e.g., due to model misspecification). We demonstrate the effectiveness of our method in applied modeling problems, ranging from disease outbreak dynamics and global warming thresholds to human decision-making. Our results support sensitivity-aware inference as a default choice for amortized Bayesian workflows, automatically providing modelers with insights into otherwise hidden dimensions.

1707. Combining Induction and Transduction for Abstract Reasoning

链接：https://iclr.cc/virtual/2025/poster/29457 abstract： When learning an input-output mapping from very few examples, is it better to first infer a latent function that explains the examples, or is it better to directly predict new test outputs, e.g. using a neural network? We study this question on ARC by training neural models for \emph{induction} (inferring latent functions) and \emph{transduction} (directly predicting the test output for a given test input). We train on synthetically generated variations of Python programs that solve ARC training tasks. We find inductive and transductive models solve different kinds of test problems, despite having the same training problems and sharing the same neural architecture: Inductive program synthesis excels at precise computations, and at composing multiple concepts, while transduction succeeds on fuzzier perceptual concepts. Ensembling them approaches human-level performance on ARC.

1708. How DNNs break the Curse of Dimensionality: Compositionality and Symmetry Learning

链接：https://iclr.cc/virtual/2025/poster/29440 abstract： We show that deep neural networks (DNNs) can efficiently learn anycomposition of functions with bounded F_{1}-norm, which allowsDNNs to break the curse of dimensionality in ways that shallow networkscannot. More specifically, we derive a generalization bound that combinesa covering number argument for compositionality, and the F_{1}-norm(or the related Barron norm) for large width adaptivity. We show thatthe global minimizer of the regularized loss of DNNs can fit for examplethe composition of two functions $f^{*}=h\circ g$ from a small numberof observations, assuming g is smooth/regular and reduces the dimensionality(e.g. g could be the quotient map of the symmetries of f^{*}),so that h can be learned in spite of its low regularity. The measuresof regularity we consider is the Sobolev norm with different levelsof differentiability, which is well adapted to the F_{1} norm.We compute scaling laws empirically and observe phase transitionsdepending on whether g or h is harder to learn, as predictedby our theory.

1709. LLM Unlearning via Loss Adjustment with Only Forget Data

链接：https://iclr.cc/virtual/2025/poster/30903 abstract： Unlearning in Large Language Models (LLMs) is essential for ensuring ethical and responsible AI use, especially in addressing privacy leak, bias, safety, and evolving regulations. Existing approaches to LLM unlearning often rely on retain data or a reference LLM, yet they struggle to adequately balance unlearning performance with overall model utility. This challenge arises because leveraging explicit retain data or implicit knowledge of retain data from a reference LLM to fine-tune the model tends to blur the boundaries between the forgotten and retain data, as different queries often elicit similar responses. In this work, we propose eliminating the need to retain data or the reference LLM for response calibration in LLM unlearning. Recognizing that directly applying gradient ascent on the forget data often leads to optimization instability and poor performance, our method guides the LLM on what not to respond to, and importantly, how to respond, based on the forget data. Hence, we introduce Forget data only Loss AjustmenT (FLAT), a "flat" loss adjustment approach which addresses these issues by maximizing f-divergence between the available template answer and the forget answer only w.r.t. the forget data. The variational form of the defined f-divergence theoretically provides a way of loss adjustment by assigning different importance weights for the learning w.r.t. template responses and the forgetting of responses subject to unlearning. Empirical results demonstrate that our approach not only achieves superior unlearning performance compared to existing methods but also minimizes the impact on the model’s retained capabilities, ensuring high utility across diverse tasks, including copyrighted content unlearning on Harry Potter dataset and MUSE Benchmark, and entity unlearning on the TOFU dataset.

1710. Topograph: An Efficient Graph-Based Framework for Strictly Topology Preserving Image Segmentation

链接：https://iclr.cc/virtual/2025/poster/29716 abstract： Topological correctness plays a critical role in many image segmentation tasks, yet most networks are trained using pixel-wise loss functions, such as Dice, neglecting topological accuracy. Existing topology-aware methods often lack robust topological guarantees, are limited to specific use cases, or impose high computational costs. In this work, we propose a novel, graph-based framework for topologically accurate image segmentation that is both computationally efficient and generally applicable. Our method constructs a component graph that fully encodes the topological information of both the prediction and ground truth, allowing us to efficiently identify topologically critical regions and aggregate a loss based on local neighborhood information. Furthermore, we introduce a strict topological metric capturing the homotopy equivalence between the union and intersection of prediction-label pairs. We formally prove the topological guarantees of our approach and empirically validate its effectiveness on binary and multi-class datasets, demonstrating state-of-the-art performance with up to fivefold faster loss computation compared to persistent homology methods.

1711. Scale-aware Recognition in Satellite Images under Resource Constraints

链接：https://iclr.cc/virtual/2025/poster/29697 abstract： Recognition of features in satellite imagery (forests, swimming pools, etc.) depends strongly on the spatial scale of the concept and therefore the resolution of the images. This poses two challenges:Which resolution is best suited for recognizing a given concept, and where and when should the costlier higher-resolution (HR) imagery be acquired? We present a novel scheme to address these challenges by introducing three components: (1) A technique to distill knowledge from models trained on HR imagery to recognition models that operate on imagery of lower resolution (LR), (2) a sampling strategy for HR imagery based on model disagreement, and (3) an LLM-based approach for inferring concept "scale". With these components we present a system to efficiently perform scale-aware recognition in satellite imagery, improving accuracy over single-scale inference while following budget constraints. Our novel approach offers up to a 26.3\% improvement over entirely HR baselines, using 76.3 \% fewer HR images. Resources are available at https://www.cs.cornell.edu/~revankar/scale_aware.

1712. KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA

链接：https://iclr.cc/virtual/2025/poster/32054 abstract： Biomedical reasoning integrates structured, codified knowledge with tacit, experience-driven insights. Depending on the context, quantity, and nature of available evidence, researchers and clinicians use diverse strategies, including rule-based, prototype-based, and case-based reasoning. Effective medical AI models must handle this complexity while ensuring reliability and adaptability. We introduce KGARevion, a knowledge graph-based agent that answers knowledge-intensive questions. Upon receiving a query, KGARevion generates relevant triplets by leveraging the latent knowledge embedded in a large language model. It then verifies these triplets against a grounded knowledge graph, filtering out errors and retaining only accurate, contextually relevant information for the final answer. This multi-step process strengthens reasoning, adapts to different models of medical inference, and outperforms retrieval-augmented generation-based approaches that lack effective verification mechanisms. Evaluations on medical QA benchmarks show that KGARevion improves accuracy by over 5.2% over 15 models in handling complex medical queries. To further assess its effectiveness, we curated three new medical QA datasets with varying levels of semantic complexity, where KGARevion improved accuracy by 10.4%. The agent integrates with different LLMs and biomedical knowledge graphs for broad applicability across knowledge-intensive tasks. We evaluated KGARevion on AfriMed-QA, a newly introduced dataset focused on African healthcare, demonstrating its strong zero-shot generalization to underrepresented medical contexts.

1713. Measuring memorization in RLHF for code completion

链接：https://iclr.cc/virtual/2025/poster/29527 abstract： Reinforcement learning with human feedback (RLHF) has become the dominant method to align large models to user preferences.Unlike fine-tuning, for which there are many studies regarding training data memorization, it is not clear how memorization is affected by or introduced in the RLHF alignment process.Understanding this relationship is important as real user data may be collected and used to align large models; if user data is memorized during RLHF and later regurgitated, this could raise privacy concerns. In addition to RLHF, other methods such as Direct Preference Optimization (DPO) and ΨPO have gained popularity for learning directly from human preferences, removing the need for optimizing intermediary reward models with reinforcement learning.In this work, we analyze how training data memorization can surface and propagate through each phase of RLHF and direct preference learning.We focus our study on code completion models, as code completion is one of the most popular use cases for large language models. We find that RLHF significantly decreases the chance that data used for reward modeling and reinforcement learning is memorized in comparison to directly fine-tuning on this data, but that examples already memorized during the fine-tuning stage of RLHF, will, in the majority of cases, remain memorized after RLHF. In contrast, we find that aligning by learning directly from human preference data via a special case of ΨPO, Identity Preference Optimization (IPO), increases the likelihood that training data is regurgitated compared to RLHF. Our work suggests that RLHF, as opposed to direct preference learning, is a safer way to mitigate the risk of regurgitating sensitive preference data when aligning large language models. We find our conclusions are robust across multiple code completion datasets, tasks, and model scales.

1714. Denoising Levy Probabilistic Models

链接：https://iclr.cc/virtual/2025/poster/29595 abstract： Investigating noise distributions beyond Gaussian in diffusion generative models remains an open challenge. The Gaussian case has been a large success experimentally and theoretically, admitting a unified stochastic differential equation (SDE) framework, encompassing score-based and denoising formulations. Recent studies have investigated the potential of \emph{heavy-tailed} noise distributions to mitigate mode collapse and effectively manage datasets exhibiting class imbalance, heavy tails, or prominent outliers. Very recently, Yoon et al.\ (NeurIPS 2023), presented the Levy-Ito model (LIM), directly extending the SDE-based framework to a class of heavy-tailed SDEs, where the injected noise followed an α-stable distribution -- a rich class of heavy-tailed distributions. Despite its theoretical elegance and performance improvements, LIM relies on highly involved mathematical techniques, which may limit its accessibility and hinder its broader adoption and further development. In this study, we take a step back, and instead of starting from the SDE formulation, we extend the denoising diffusion probabilistic model (DDPM) by directly replacing the Gaussian noise with α-stable noise. By using only elementary proof techniques, we show that the proposed approach, \emph{denoising L\'{e}vy probabilistic model} (DLPM) algorithmically boils down to running vanilla DDPM with minor modifications, hence allowing the use of existing implementations with minimal changes. Remarkably, as opposed to the Gaussian case, DLPM and LIM yield different training algorithms and different backward processes, leading to distinct sampling algorithms. This fundamental difference translates favorably for the performance of DLPM in various aspects: our experiments show that DLPM achieves better coverage of the tails of the data distribution, better generation of unbalanced datasets, and improved computation times requiring significantly smaller number of backward steps.

1715. Why In-Context Learning Models are Good Few-Shot Learners?

链接：https://iclr.cc/virtual/2025/poster/28701 abstract： We explore in-context learning (ICL) models from a learning-to-learn perspective. Unlike studies that identify specific learning algorithms in ICL models, we compare ICL models with typical meta-learners to understand their superior performance. We theoretically prove the expressiveness of ICL models as learning algorithms and examine their learnability and generalizability. Our findings show that ICL with transformers can effectively construct data-dependent learning algorithms instead of directly follow existing ones (including gradient-based, metric-based, and amortization-based meta-learners). The construction of such learning algorithm is determined by the pre-training process, as a function fitting the training distribution, which raises generalizability as an important issue.With above understanding, we propose strategies to transfer techniques for classical deep networks to meta-level to further improve ICL. As examples, we implement meta-level meta-learning for domain adaptability with limited data and meta-level curriculum learning for accelerated convergence during pre-training, demonstrating their empirical effectiveness.

1716. Boltzmann-Aligned Inverse Folding Model as a Predictor of Mutational Effects on Protein-Protein Interactions

链接：https://iclr.cc/virtual/2025/poster/28490 abstract： Predicting the change in binding free energy ($\Delta \Delta G$) is crucial for understanding and modulating protein-protein interactions, which are critical in drug design.Due to the scarcity of experimental $\Delta\Delta G$ data, existing methods focus on pre-training, while neglecting the importance of alignment.In this work, we propose Boltzmann Alignment technique to transfer knowledge from pre-trained inverse folding models to prediction of $\Delta\Delta G$.We begin by analyzing the thermodynamic definition of $\Delta\Delta G$ and introducing the Boltzmann distribution to connect energy to the protein conformational distribution. However, the protein conformational distribution is intractable. Therefore, we employ Bayes' theorem to circumvent direct estimation and instead utilize the log-likelihood provided by protein inverse folding models for the estimation of $\Delta\Delta G$. Compared to previous methods based on inverse folding, our method explicitly accounts for the unbound state of the protein complex in the $\Delta \Delta G$ thermodynamic cycle, introducing a physical inductive bias and achieving supervised and unsupervised state-of-the-art (SoTA) performance.Experimental results on SKEMPI v2 indicate that our method achieves Spearman coefficients of 0.3201

(unsupervised) and 0.5134 (supervised) on SKEMPI v2, significantly surpassing the previously reported %SoTA valuesSoTA results of 0.2632 and 0.4324, respectively.Furthermore, we demonstrate the capability of our method in bindingenergy prediction, protein-protein docking, and antibody optimization tasks.Code is available at https://github.com/aim-uofa/BA-DDG

1717. Co$^{\mathbf{3}}$Gesture: Towards Coherent Concurrent Co-speech 3D Gesture Generation with Interactive Diffusion

链接：https://iclr.cc/virtual/2025/poster/29405 abstract： Generating gestures from human speech has gained tremendous progress in animating virtual avatars. While the existing methods enable synthesizing gestures cooperated by people self-talking, they overlook the practicality of concurrent gesture modeling with two-person interactive conversations. Moreover, the lack of high-quality datasets with concurrent co-speech gestures also limits handling this issue. To fulfill this goal, we first construct a large-scale concurrent co-speech gesture dataset that contains more than 7M frames for diverse two-person interactive posture sequences, dubbed $\textbf{GES-Inter}$. Moreover, we propose Co$^{\mathbf{3}}$Gesture, a novel framework that enables concurrent coherent co-speech gesture synthesis including two-person interactive movements. Our framework is built upon two cooperative generation branches conditioned on decomposed speaker audio. Specifically, to enhance the coordination of human postures w.r.t corresponding speaker audios while interacting with the conversational partner, we present a Temporal-Interaction Module ($\textbf{TIM}$). TIM can effectively model the temporal association representation between two speakers' gesture sequences as interaction guidance and fuse it into the concurrent gesture generation. Then, we devise a mutual attention mechanism to further boost learning dependencies of interacted concurrent motions, thereby enabling us to generate vivid and coherent gestures. Extensive experiments demonstrate that our method outperforms the state-of-the-art models on our newly collected GES-Inter dataset.

1718. Latent Action Pretraining from Videos

链接：https://iclr.cc/virtual/2025/poster/29409 abstract： We introduce Latent Action Pretraining for general Action models (LAPA), the first unsupervised method for pretraining Vision-Language-Action (VLA) models without ground-truth robot action labels. Existing Vision-Language-Action models require action labels typically collected by human teleoperators during pretraining, which significantly limits possible data sources and scale. In this work, we propose a method to learn from internet-scale videos that do not have robot action labels. We first train an action quantization model leveraging VQ-VAE-based objective to learn discrete latent actions between image frames, then pretrain a latent VLA model to predict these latent actions from observations and task descriptions, and finally finetune the VLA on small-scale robot manipulation data to map from latent to robot actions. Experimental results demonstrate that our method significantly outperforms existing techniques that train robot manipulation policies from large-scale videos. Furthermore, it outperforms the state-of-the-art VLA model trained with robotic action labels on real-world manipulation tasks that require language conditioning, generalization to unseen objects, and semantic generalization to unseen instructions. Training only on human manipulation videos also shows positive transfer, opening up the potential for leveraging web-scale data for robotics foundation models.

1719. Improving Generalization and Robustness in SNNs Through Signed Rate Encoding and Sparse Encoding Attacks

链接：https://iclr.cc/virtual/2025/poster/28253 abstract： Rate-encoded spiking neural networks (SNNs) are known to offer superior adversarial robustness compared to direct-encoded SNNs but have relatively poor generalization on clean input. While the latter offers good generalization on clean input it suffers poor adversarial robustness under standard training. A key reason for this difference is the input noise introduced by the rate encoding, which encodes a pixel intensity with T independent Bernoulli samples. To improve the generalization of rate-encoded SNNs, we propose the *signed rate encoding* (sRATE) that allows mean centering of the input and helps reduce the randomness introduced by the encoding, resulting in improved clean accuracy. In contrast to rate encoding, where input restricted to $[0,1]^d$ is encoded in $\{0,1\}^{d\times T}$, the signed rate encoding allows input in $[-1,1]^d$ to be encoded with spikes in $\{-1,0,1\}^{d\times T}$, where positive (negative) inputs are encoded with positive (negative) spikes. We further construct efficient \textit{Sparse Encoding Attack} (SEA) on standard and signed rate encoded input, which performs l_0-norm restricted adversarial attack in the discrete encoding space. We prove the theoretical optimality of the attack under the first-order approximation of the loss and compare it empirically with the existing attacks on the input space. Adversarial training performed with SEA, under signed rate encoding, offers superior adversarial robustness to the existing attacks and itself. Experiments conducted on standard datasets show the effectiveness of sign rate encoding in improving accuracy across all settings including adversarial robustness. The code is available at https://github.com/BhaskarMukhoty/SignedRateEncoding

1720. Event-Driven Online Vertical Federated Learning

链接：https://iclr.cc/virtual/2025/poster/30357 abstract： Online learning is more adaptable to real-world scenarios in Vertical Federated Learning (VFL) compared to offline learning. However, integrating online learning into VFL presents challenges due to the unique nature of VFL, where clients possess non-intersecting feature sets for the same sample. In real-world scenarios, the clients may not receive data streaming for the disjoint features for the same entity synchronously. Instead, the data are typically generated by an event relevant to only a subset of clients.We are the first to identify these challenges in online VFL, which have been overlooked by previous research. To address these challenges, we proposed an event-driven online VFL framework. In this framework, only a subset of clients were activated during each event, while the remaining clients passively

collaborated in the learning process. Furthermore, we incorporated dynamic local regret (DLR) into VFL to address the challenges posed by online learning problems with non-convex models within a non-stationary environment.We conducted a comprehensive regret analysis of our proposed framework, specifically examining the DLR under non-convex conditions with event-driven online VFL. Extensive experiments demonstrated that our proposed framework was more stable than the existing online VFL framework under non-stationary data conditions while also significantly reducing communication and computation costs.

1721. Do WGANs succeed because they minimize the Wasserstein Distance? Lessons from Discrete Generators

链接：https://iclr.cc/virtual/2025/poster/30814 abstract： Since WGANs were first introduced, there has been considerable debate whether their success in generating realistic images can be attributed to minimizing the Wasserstein distance between the distribution of generated images and the training distribution. In this paper we present theoretical and experimental results that show that successful WGANs {\em do} minimize the Wasserstein distance but the form of the distance that is minimized depends highly on the discriminator architecture and its inductive biases. Specifically, we show that when the discriminator is convolutional, WGANs minimize the Wasserstein distance between {\em patches} in the generated images and the training images, not the Wasserstein distance between images.Our results are obtained by considering {\em discrete} generators for which the Wasserstein distance between the generator distribution and the training distribution can be computed exactly and the minimum can be characterized analytically. We present experimental results with discrete GANs that generate realistic fake images (comparable in quality to their continuous counterparts) and present evidence that they are minimizing the Wasserstein distance between real and fake patches and not the distance between real and fake images.

1722. Towards Robust and Parameter-Efficient Knowledge Unlearning for LLMs

链接：https://iclr.cc/virtual/2025/poster/31216 abstract： Large Language Models (LLMs) have demonstrated strong reasoning and memorization capabilities via pretraining on massive textual corpora. However, this poses risk of privacy and copyright violations, highlighting the need for efficient machine unlearning methods that remove sensitive data without retraining from scratch. While Gradient Ascent (GA) is commonly used to unlearn by reducing the likelihood of generating unwanted content, it leads to unstable optimization and catastrophic forgetting of retrained knowledge. We find that combining GA with low-rank adaptation results in poor trade-offs between computational cost and generative performance. To address these challenges, we propose Low-rank Knowledge Unlearning (LoKU), a novel framework that enables robust and efficient unlearning for LLMs. First, we introduce Inverted Hinge Loss, which suppresses unwanted tokens while maintaining fluency by boosting the probability of the next most likely token. Second, we develop a data-adaptive initialization for LoRA adapters via low-rank approximation weighted with relative Fisher information, thereby focusing updates on parameters critical for removing targeted knowledge. Experiments on the Training Data Extraction Challenge dataset using GPT-Neo models as well as on the TOFU benchmark with Phi-1.5B and Llama2-7B models demonstrate that our approach effectively removes sensitive information while maintaining reasoning and generative capabilities with minimal impact. Our implementation can be found in https://github.com/csm9493/efficient-llm-unlearning.

1723. Training Free Exponential Context Extension via Cascading KV Cache

链接：https://iclr.cc/virtual/2025/poster/28984 abstract： The transformer's context window is vital for tasks such as few-shot learning and conditional generation as it preserves previous tokens for active memory. However, as the context lengths increase, the computational costs grow quadratically, hindering the deployment of large language models (LLMs) in real-world, long sequence scenarios. Although some recent key-value caching (KV Cache) methods offer linear inference complexity, they naively manage the stored context, prematurely evicting tokens and losing valuable information. Moreover, they lack an optimized prefill/prompt stage strategy, resulting in higher latency than even quadratic attention for realistic context sizes. In response, we introduce a novel mechanism that leverages cascading sub-cache buffers to selectively retain the most relevant tokens, enabling the model to maintain longer context histories without increasing the cache size. Our approach outperforms linear caching baselines across key benchmarks, including streaming perplexity, question answering, book summarization, and passkey retrieval, where it retains better retrieval accuracy at 1M tokens after four doublings of the cache size of 65K. Additionally, our method reduces prefill stage latency by a factor of 6.8 when compared to flash attention on 1M tokens. These innovations not only enhance the computational efficiency of LLMs but also pave the way for their effective deployment in resource-constrained environments, enabling large-scale, real-time applications with significantly reduced latency.

1724. Mixture Compressor for Mixture-of-Experts LLMs Gains More

链接：https://iclr.cc/virtual/2025/poster/28749 abstract： Mixture-of-Experts large language models (MoE-LLMs) marks a significant step forward of language models, however, they encounter two critical challenges in practice: 1) expert parameters lead to considerable memory consumption and loading latency; and 2) the current activated experts are redundant, as many tokens may only require a single expert. Motivated by these issues, we investigate the MoE-LLMs and make two key observations: a) different experts exhibit varying behaviors on activation reconstruction error, routing scores, and activated frequencies, highlighting their differing importance, and b) not all tokens are equally important-- only a small subset is critical. Building on these insights, we propose MC, a training-free Mixture-Compressor for MoE-LLMs, which leverages the significance

of both experts and tokens to achieve an extreme compression. First, to mitigate storage and loading overheads, we introduce Pre-Loading Mixed-Precision Quantization (PMQ), which formulates the adaptive bit-width allocation as a Linear Programming (LP) problem, where the objective function balances multi-factors reflecting the importance of each expert. Additionally, we develop Online Dynamic Pruning (ODP), which identifies important tokens to retain and dynamically select activated experts for other tokens during inference to optimize efficiency while maintaining performance. Our MC integrates static quantization and dynamic pruning to collaboratively achieve extreme compression for MoE-LLMs with less accuracy loss, ensuring an optimal trade-off between performance and efficiency Extensive experiments confirm the effectiveness of our approach. For instance, at 2.54 bits, MC compresses 76.6% of the model, with only a 3.8% average accuracy loss. During dynamic inference, we further reduce activated parameters by 15%, with a performance drop of less than 0.6%. Remarkably, MC even surpasses floating-point 13b dense LLMs with significantly smaller parameter sizes, suggesting that mixture compression in MoE-LLMs has the potential to outperform both comparable and larger dense LLMs. Our code isavailable at https://github.com/Aaronhuang-778/MC-MoE

1725. A Training-Free Sub-quadratic Cost Transformer Model Serving Framework with Hierarchically Pruned Attention

链接：https://iclr.cc/virtual/2025/poster/29752 abstract： In modern large language models (LLMs), increasing the context length is crucial for improving comprehension and coherence in long-context, multi-modal, and retrieval-augmented language generation. While many recent transformer models attempt to extend their context length over a million tokens, they remain impractical due to the quadratic time and space complexities.Although recent works on linear and sparse attention mechanisms can achieve this goal, their real-world applicability is often limited by the need to re-train from scratch and significantly worse performance. In response, we propose a novel approach, Hierarchically Pruned Attention (HiP), which reduces the time complexity of the attention mechanism to $O(T \log T)$ and the space complexity to $O(T)$, where T is the sequence length. We notice a pattern in the attention scores of pretrained LLMs where tokens close together tend to have similar scores, which we call "attention locality". Based on this observation, we utilize a novel tree-search-like algorithm that estimates the top-k key tokens for a given query on the fly, which is mathematically guaranteed to have better performance than random attention pruning. In addition to improving the time complexity of the attention mechanism, we further optimize GPU memory usage by implementing KV cache offloading, which stores only $O(\log T)$ tokens on the GPU while maintaining similar decoding throughput. Experiments on benchmarks show that HiP, with its training-free nature, significantly reduces both prefill and decoding latencies, as well as memory usage, while maintaining high-quality generation with minimal degradation.HiP enables pretrained LLMs to scale up to millions of tokens on commodity GPUs, potentially unlocking long-context LLM applications previously deemed infeasible.

1726. Transformers Handle Endogeneity in In-Context Linear Regression

链接：https://iclr.cc/virtual/2025/poster/29683 abstract： We explore the capability of transformers to address endogeneity in in-context linear regression. Our main finding is that transformers inherently possess a mechanism to handle endogeneity effectively using instrumental variables (IV). First, we demonstrate that the transformer architecture can emulate a gradient-based bi-level optimization procedure that converges to the widely used two-stage least squares (2SLS) solution at an exponential rate. Next, we propose an in-context pretraining scheme and provide theoretical guarantees showing that the global minimizer of the pre-training loss achieves a small excess loss. Our extensive experiments validate these theoretical findings, showing that the trained transformer provides more robust and reliable in-context predictions and coefficient estimates than the 2SLS method, in the presence of endogeneity.

1727. Language-Assisted Feature Transformation for Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/31115 abstract： This paper introduces LAFT, a novel feature transformation method designed to incorporate user knowledge and preferences into anomaly detection using natural language. Accurately modeling the boundary of normality is crucial for distinguishing abnormal data, but this is often challenging due to limited data or the presence of nuisance attributes. While unsupervised methods that rely solely on data without user guidance are common, they may fail to detect anomalies of specific interest. To address this limitation, we propose Language-Assisted Feature Transformation (LAFT), which leverages the shared image-text embedding space of vision-language models to transform visual features according to user-defined requirements. Combined with anomaly detection methods, LAFT effectively aligns visual features with user preferences, allowing anomalies of interest to be detected. Extensive experiments on both toy and real-world datasets validate the effectiveness of our method.

1728. Training Language Models on Synthetic Edit Sequences Improves Code Synthesis

链接：https://iclr.cc/virtual/2025/poster/30611 abstract： Software engineers mainly write code by editing existing programs. In contrast, language models (LMs) autoregressively synthesize programs in a single pass. One explanation for this is the scarcity of sequential edit data. While high-quality instruction data for code synthesis is scarce, edit data for synthesis is even scarcer. To fill this gap, we develop a synthetic data generation algorithm called LintSeq. This algorithm refactors programs into sequences of synthetic edits by using a linter to procedurally sample across interdependent lines of source code. Synthetic edits sampled with LintSeq reflect the syntax and semantics of their programming language. To test the algorithm, we use it to refactor

a dataset of instruction + program pairs into instruction + program-diff-sequence tuples. Then, we fine-tune a series of smaller LMs ranging from 2.6B to 14B parameters on both the re-factored and original versions of this dataset. We perform comprehensive evaluations comparing edit sequence code LMs against baselines on HumanEval, MBPP(+), CodeContests, DS-1000, and BigCodeBench. We show that models fine-tuned to iteratively synthesize code match or outperform baselines on pass@1, and exhibit better scaling across higher pass@k as a function of total test-time FLOPs. Finally, we also pretrain our own tiny LMs for code understanding. We show that fine-tuning these models to synthesize code edit-by-edit results in strong performance on HumanEval and MBPP(+) compared to existing code language models of similar scale such as CodeT5+, AlphaCode, and Codex.

1729. MOFFlow: Flow Matching for Structure Prediction of Metal-Organic Frameworks

链接：https://iclr.cc/virtual/2025/poster/28990 abstract： Metal-organic frameworks (MOFs) are a class of crystalline materials with promising applications in many areas such as carbon capture and drug delivery. In this work, we introduce MOFFlow, the first deep generative model tailored for MOF structure prediction. Existing approaches, including ab initio calculations and even deep generative models, struggle with the complexity of MOF structures due to the large number of atoms in the unit cells. To address this limitation, we propose a novel Riemannian flow matching framework that reduces the dimensionality of the problem by treating the metal nodes and organic linkers as rigid bodies, capitalizing on the inherent modularity of MOFs. By operating in the $SE(3)$ space, MOFFlow effectively captures the roto-translational dynamics of these rigid components in a scalable way. Our experiment demonstrates that MOFFlow accurately predicts MOF structures containing several hundred atoms, significantly outperforming conventional methods and state-of-the-art machine learning baselines while being much faster. Code available at https://github.com/nayoung10/MOFFlow.

1730. BALROG: Benchmarking Agentic LLM and VLM Reasoning On Games

链接：https://iclr.cc/virtual/2025/poster/28856 abstract： Large Language Models (LLMs) and Vision Language Models (VLMs) possess extensive knowledge and exhibit promising reasoning abilities, however, they still struggle to perform well in complex, dynamic environments. Real-world tasks require handling intricate interactions, advanced spatial reasoning, long-term planning, and continuous exploration of new strategies—areas in which we lack effective methodologies for comprehensively evaluating these capabilities. To address this gap, we introduce BALROG, a novel benchmark designed to assess the agentic capabilities of LLMs and VLMs through a diverse set of challenging games. Our benchmark incorporates a range of existing reinforcement learning environments with varying levels of difficulty, including tasks that are solvable by non-expert humans in seconds to extremely challenging ones that may take years to master (e.g., the NetHack Learning Environment). We devise fine-grained metrics to measure performance and conduct an extensive evaluation of several popular open-source and closed-source LLMs and VLMs. Our findings indicate that while current models achieve partial success in the easier games, they struggle significantly with more challenging tasks. Notably, we observe severe deficiencies in vision-based decision-making, as several models perform worse when visual representations of the environments are provided. We release BALROG as an open and user-friendly benchmark to facilitate future research and development in the agentic community. Code and Leaderboard at balrogai.com

1731. The Illustrated AlphaFold

链接：https://iclr.cc/virtual/2025/poster/31362 abstract： We present the Illustrated AlphaFold, a visual walkthrough of the architecture and information flow of AlphaFold 3. We explain every model component and training detail, with particular focus on the advances since AlphaFold 2 – including the unified tokenization scheme that extends to DNA, RNA, and small molecules, as well as the novel diffusion-based structural module. Finally, we include some musings on the ML lessons learned from studying AlphaFold 3.

1732. Noise-conditioned Energy-based Annealed Rewards (NEAR): A Generative Framework for Imitation Learning from Observation

链接：https://iclr.cc/virtual/2025/poster/30464 abstract： This paper introduces a new imitation learning framework based on energy-based generative models capable of learning complex, physics-dependent, robot motion policies through state-only expert motion trajectories. Our algorithm, called Noise-conditioned Energy-based Annealed Rewards (NEAR), constructs several perturbed versions of the expert's motion data distribution and learns smooth, and well-defined representations of the data distribution's energy function using denoising score matching. We propose to use these learnt energy functions as reward functions to learn imitation policies via reinforcement learning. We also present a strategy to gradually switch between the learnt energy functions, ensuring that the learnt rewards are always well-defined in the manifold of policy-generated samples. We evaluate our algorithm on complex humanoid tasks such as locomotion and martial arts and compare it with state-only adversarial imitation learning algorithms like Adversarial Motion Priors (AMP). Our framework sidesteps the optimisation challenges of adversarial imitation learning techniques and produces results comparable to AMP in several quantitative metrics across multiple imitation settings.

1733. Improved Convergence Rate for Diffusion Probabilistic Models

链接：https://iclr.cc/virtual/2025/poster/29610 abstract： Score-based diffusion models have achieved remarkable empirical performance in the field of machine learning and artificial intelligence for their ability to generate high-quality new data instances from complex distributions. Improving our understanding of diffusion models, including mainly convergence analysis for such models, has attracted a lot of interests. Despite a lot of theoretical attempts, there still exists significant gap between theory and practice. Towards to close this gap, we establish an iteration complexity at the order of $d^{1/3}\varepsilon^{-2/3}$, which is better than $d^{5/12}\varepsilon^{-1}$, the best known complexity achieved before our work. This convergence analysis is based on a randomized midpoint method, which is first proposed for log-concave sampling (Shen & Lee, 2019), and then extended to diffusion models by Gupta et al. (2024). Our theory accommodates ε-accurate score estimates, and does not require log-concavity on the target distribution. Moreover, the algorithm can also be parallelized to run in only $O(\log^2(d/\varepsilon))$ parallel rounds in a similar way to prior works.

1734. Probing the Latent Hierarchical Structure of Data via Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/31261 abstract：

1735. SymDiff: Equivariant Diffusion via Stochastic Symmetrisation

链接：https://iclr.cc/virtual/2025/poster/28724 abstract： We propose SymDiff, a method for constructing equivariant diffusion models using the framework of stochastic symmetrisation. SymDiff resembles a learned data augmentation that is deployed at sampling time, and is lightweight, computationally efficient, and easy to implement on top of arbitrary off-the-shelf models. In contrast to previous work, SymDiff typically does not require any neural network components that are intrinsically equivariant, avoiding the need for complex parameterisations or the use of higher-order geometric features. Instead, our method can leverage highly scalable modern architectures as drop-in replacements for these more constrained alternatives. We show that this additional flexibility yields significant empirical benefit for E(3)-equivariant molecular generation. To the best of our knowledge, this is the first application of symmetrisation to generative modelling, suggesting its potential in this domain more generally.

1736. Prediction Risk and Estimation Risk of the Ridgeless Least Squares Estimator under General Assumptions on Regression Errors

链接：https://iclr.cc/virtual/2025/poster/30610 abstract： In recent years, there has been a significant growth in research focusing on minimum ℓ_2 norm (ridgeless) interpolation least squares estimators. However, the majority of these analyses have been limited to an unrealistic regression error structure, assuming independent and identically distributed errors with zero mean and common variance. In this paper, we explore prediction risk as well as estimation risk under more general regression error assumptions, highlighting the benefits of overparameterization in a more realistic setting that allows for clustered or serial dependence. Notably, we establish that the estimation difficulties associated with the variance components of both risks can be summarized through the trace of the variance-covariance matrix of the regression errors. Our findings suggest that the benefits of overparameterization can extend to time series, panel and grouped data.

1737. L3Ms — Lagrange Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29479 abstract： Supervised fine-tuning (SFT) and alignment of large language models (LLMs) are key steps in providing a good user experience. However, the concept of an appropriate alignment is inherently application-dependent, and current methods often rely on heuristic choices to drive optimization. In this work, we formulate SFT and alignment as a constrained optimization problem: the LLM is fine-tuned on a task while being required to meet application-specific requirements, without resorting to heuristics. To solve this, we propose Lagrange Large Language Models (L3Ms), which employ logarithmic barriers to enforce the constraints. This approach allows for the customization of L3Ms across diverse applications while avoiding heuristic-driven processes. We experimentally demonstrate the versatility and efficacy of L3Ms in achieving tailored alignments for various applications.

1738. Learning High-Degree Parities: The Crucial Role of the Initialization

链接：https://iclr.cc/virtual/2025/poster/29795 abstract： Parities have become a standard benchmark for evaluating learning algorithms. Recent works show that regular neural networks trained by gradient descent can efficiently learn degree k parities on uniform inputs for constant k, but fail to do so when k and $d-k$ grow with d (here d is the ambient dimension). However, the case where $k=d-O_d(1)$, including the degree d parity (the full parity), has remained unsettled. This paper shows that for gradient descent on regular neural networks, learnability depends on the initial weight distribution. On one hand, the discrete Rademacher initialization enables efficient learning of almost-full parities, while on the other hand, its Gaussian perturbation with large enough constant standard deviation σ prevents it. The positive result for almost-full parities is shown to hold up to $\sigma=O(d^{-1})$, pointing to questions about a sharper threshold phenomenon. Unlike statistical query (SQ) learning, where a singleton function class like the full parity is trivially learnable, our negative result applies to a fixed function and relies on an initial gradient alignment}measure of potential broader relevance to neural networks learning.

1739. Learning to Contextualize Web Pages for Enhanced Decision Making

by LLM Agents

链接：https://iclr.cc/virtual/2025/poster/31086 abstract： Recent advances in large language models (LLMs) have led to a growing interest in developing LLM-based agents for automating web tasks. However, these agents often struggle with even simple tasks on real-world websites due to their limited capability to understand and process complex web page structures. In this work, we introduce LCoW, a framework for Learning language models to Contextualize complex Web pages into a more comprehensible form, thereby enhancing decision making by LLM agents. LCoW decouples web page understanding from decision making by training a separate contextualization module to transform complex web pages into comprehensible format, which are then utilized by the decision-making agent. We demonstrate that our contextualization module effectively integrates with LLM agents of various scales to significantly enhance their decision-making capabilities in web automation tasks. Notably, LCoW improves the success rates of closed-source LLMs (e.g., Gemini-1.5-flash, GPT-4o, Claude-3.5-Sonnet) by an average of 15.6%, and demonstrates a 23.7% average improvement in success rates for open-source LMs (e.g., Llama-3.1-8B, Llama-3.1-70B) on the WorkArena benchmark.Moreover, the Gemini-1.5-flash agent with LCoW achieves state-of-the-art results on the WebShop benchmark, outperforming human experts. The relevant code materials are available at our project page: https://lcowiclr2025.github.io.

1740. Microcanonical Langevin Ensembles: Advancing the Sampling of Bayesian Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29695 abstract： Despite recent advances, sampling-based inference for Bayesian Neural Networks (BNNs) remains a significant challenge in probabilistic deep learning. While sampling-based approaches do not require a variational distribution assumption, current state-of-the-art samplers still struggle to navigate the complex and highly multimodal posteriors of BNNs. As a consequence, sampling still requires considerably longer inference times than non-Bayesian methods even for small neural networks, despite recent advances in making software implementations more efficient. Besides the difficulty of finding high-probability regions, the time until samplers provide sufficient exploration of these areas remains unpredictable. To tackle these challenges, we introduce an ensembling approach that leverages strategies from optimization and a recently proposed sampler called Microcanonical Langevin Monte Carlo (MCLMC) for efficient, robust and predictable sampling performance. Compared to approaches based on the state-of-the-art No-U-Turn Sampler, our approach delivers substantial speedups up to an order of magnitude, while maintaining or improving predictive performance and uncertainty quantification across diverse tasks and data modalities. The suggested Microcanonical Langevin Ensembles and modifications to MCLMC additionally enhance the method's predictability in resource requirements, facilitating easier parallelization. All in all, the proposed method offers a promising direction for practical, scalable inference for BNNs.

1741. TIGER: Time-frequency Interleaved Gain Extraction and Reconstruction for Efficient Speech Separation

链接：https://iclr.cc/virtual/2025/poster/28155 abstract： In recent years, much speech separation research has focused primarily on improving model performance. However, for low-latency speech processing systems, high efficiency is equally important. Therefore, we propose a speech separation model with significantly reduced parameters and computational costs: Time-frequency Interleaved Gain Extraction and Reconstruction network (TIGER). TIGER leverages prior knowledge to divide frequency bands and compresses frequency information. We employ a multi-scale selective attention module to extract contextual features, while introducing a full-frequency-frame attention module to capture both temporal and frequency contextual information. Additionally, to more realistically evaluate the performance of speech separation models in complex acoustic environments, we introduce a dataset called EchoSet. This dataset includes noise and more realistic reverberation (e.g., considering object occlusions and material properties), with speech from two speakers overlapping at random proportions. Experimental results showed that models trained on EchoSet had better generalization ability than those trained on other datasets to the data collected in the physical world, which validated the practical value of the EchoSet. On EchoSet and real-world data, TIGER significantly reduces the number of parameters by 94.3% and the MACs by 95.3% while achieving performance surpassing state-of-the-art (SOTA) model TF-GridNet.

1742. A Watermark for Order-Agnostic Language Models

链接：https://iclr.cc/virtual/2025/poster/29861 abstract： Statistical watermarking techniques are well-established for sequentially decoded language models (LMs). However, these techniques cannot be directly applied to order-agnostic LMs, as the tokens in order-agnostic LMs are not generated sequentially. In this work, we introduce PATTERN-MARK, a pattern-based watermarking framework specifically designed for order-agnostic LMs. We develop aMarkov-chain-based watermark generator that produces watermark key sequences with high-frequency key patterns. Correspondingly, we propose a statistical pattern-based detection algorithm that recovers the key sequence during detection and conducts statistical tests based on the count of high-frequency patterns. Our extensive evaluations on order-agnostic LMs, such as ProteinMPNN and CMLM, demonstrate PATTERN-MARK’s enhanced detection efficiency, generation quality, and robustness, positioning it as a superior watermarking technique for order-agnostic LMs.

1743. Efficient Neuron Segmentation in Electron Microscopy by Affinity-Guided Queries

链接：https://iclr.cc/virtual/2025/poster/29280 abstract： Accurate segmentation of neurons in electron microscopy (EM) images plays a crucial role in understanding the intricate wiring patterns of the brain. Existing automatic neuron segmentation methods rely on traditional clustering algorithms, where affinities are predicted first, and then watershed and post-processing algorithms are applied to yield segmentation results. Due to the nature of watershed algorithm, this paradigm has deficiency in both prediction quality and speed. Inspired by recent advances in natural image segmentation, we propose to use query-based methods to address the problem because they do not necessitate watershed algorithms. However, we find that directly applying existing query-based methods faces great challenges due to the large memory requirement of the 3D data and considerably different morphology of neurons. To tackle these challenges, we introduce affinity-guided queries and integrate them into a lightweight query-based framework. Specifically, we first predict affinities with a lightweight branch, which provides coarse neuron structure information. The affinities are then used to construct affinity-guided queries, facilitating segmentation with bottom-up cues. These queries, along with additional learnable queries, interact with the image features to directly predict the final segmentation results. Experiments on benchmark datasets demonstrated that our method achieved better results over state-of-the-art methods with a 2$\sim$3$\times$ speedup in inference. Code is available at https://github.com/chenhang98/AGQ.

1744. ADBM: Adversarial Diffusion Bridge Model for Reliable Adversarial Purification

链接：https://iclr.cc/virtual/2025/poster/28839 abstract： Recently Diffusion-based Purification (DiffPure) has been recognized as an effective defense method against adversarial examples. However, we find DiffPure which directly employs the original pre-trained diffusion models for adversarial purification, to be suboptimal. This is due to an inherent trade-off between noise purification performance and data recovery quality. Additionally, the reliability of existing evaluations for DiffPure is questionable, as they rely on weak adaptive attacks. In this work, we propose a novel Adversarial Diffusion Bridge Model, termed ADBM. ADBM directly constructs a reverse bridge from the diffused adversarial data back to its original clean examples, enhancing the purification capabilities of the original diffusion models. Through theoretical analysis and experimental validation across various scenarios, ADBM has proven to be a superior and robust defense mechanism, offering significant promise for practical applications. Code is available at https://github.com/LixiaoTHU/ADBM.

1745. LoLCATs: On Low-Rank Linearizing of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30763 abstract： Recent works show we can linearize large language models (LLMs)—swapping the quadratic attentions of popular Transformer-based LLMs with subquadratic analogs, such as linear attention—avoiding the expensive pretraining costs. However, linearizing LLMs often significantly degrades model quality, still requires training over billions of tokens, and remains limited to smaller 1.3B to 7B LLMs. We thus propose Low-rank Linear Conversion via Attention Transfer (LoLCATs), a simple two-step method that improves LLM linearizing quality with orders of magnitudes less memory and compute. We base these steps on two findings. First, we can replace an LLM's softmax attentions with closely-approximating linear attentions, simply by *training* the linear attentions to match their softmax counterparts with an output MSE loss (“attention transfer”). Then, this enables adjusting for approximation errors and recovering LLM quality simply with *low-rank* adaptation (LoRA). LoLCATs significantly improves linearizing quality, training efficiency, and scalability. We significantly reduce the linearizing quality gap and produce state-of-the-art subquadratic LLMs from Llama 3 8B and Mistral 7B v0.1, leading to 20+ points of improvement on 5-shot MMLU. Furthermore, LoLCATs does so with only 0.2% of past methods' model parameters and 0.04-0.2% of their training tokens. Finally, we apply LoLCATs to create the first linearized 70B and 405B LLMs (50$\times$ that of prior work). When compared with prior approaches under the same compute budgets, LoLCATs significantly improves linearizing quality, closing the gap between linearized and original Llama 3.1 70B and 405B LLMs by 77.8\% and 78.1\% on 5-shot MMLU.

1746. DynAlign: Unsupervised Dynamic Taxonomy Alignment for Cross-Domain Segmentation

链接：https://iclr.cc/virtual/2025/poster/30161 abstract： Current unsupervised domain adaptation (UDA) methods for semantic segmentation typically assume identical class labels between the source and target domains. This assumption ignores the label-level domain gap, which is common in real-world scenarios, and limits their ability to identify finer-grained or novel categories without requiring extensive manual annotation.A promising direction to address this limitation lies in recent advancements in foundation models, which exhibit strong generalization abilities due to their rich prior knowledge. However, these models often struggle with domain-specific nuances and underrepresented fine-grained categories.To address these challenges, we introduce DynAlign, a two-stage framework that integrates UDA with foundation models to bridge both the image-level and label-level domain gaps. Our approach leverages prior semantic knowledge to align source categories with target categories that can be novel, more fine-grained, or named differently. (e.g., vehicle to car, truck, bus). Foundation models are then employed for precise segmentation and category reassignment. To further enhance accuracy, we propose a knowledge fusion approach that dynamically adapts to varying scene contexts. DynAlign generates accurate predictions in a new target label space without requiring any manual annotations, allowing seamless adaptation to new taxonomies through either model retraining or direct inference. Experiments on the GTA $\rightarrow$ IDD and GTA$\rightarrow$ Mapillary benchmarks validate the effectiveness of our approach, achieving a significant improvement over existing methods. Our code is publically available at https://github.com/hansunhayden/DynAlign.

1747. On Conformal Isometry of Grid Cells: Learning Distance-Preserving Position Embedding

链接：https://iclr.cc/virtual/2025/poster/29290 abstract： This paper investigates the conformal isometry hypothesis as a potential explanation for the hexagonal periodic patterns in grid cell response maps. We posit that grid cell activities form a high-dimensional vector in neural space, encoding the agent's position in 2D physical space. As the agent moves, this vector rotates within a 2D manifold in the neural space, driven by a recurrent neural network. The conformal hypothesis proposes that this neural manifold is a conformal isometric embedding of 2D physical space, where local physical distance is preserved by the embedding up to a scaling factor (or unit of metric). Such distance-preserving position embedding is indispensable for path planning in navigation, especially planning local straight path segments. We conduct numerical experiments to show that this hypothesis leads to the hexagonal grid firing patterns by learning maximally distance-preserving position embedding, agnostic to the choice of the recurrent neural network. Furthermore, we present a theoretical explanation of why hexagon periodic patterns emerge by minimizing our loss function by showing that hexagon flat torus is maximally distance preserving.

1748. Amortized Control of Continuous State Space Feynman-Kac Model for Irregular Time Series

链接：https://iclr.cc/virtual/2025/poster/30730 abstract： Many real-world datasets, such as healthcare, climate, and economics, are often collected as irregular time series, which poses challenges for accurate modeling. In this paper, we propose the Amortized Control of continuous State Space Model (ACSSM) for continuous dynamical modeling of time series for irregular and discrete observations. We first present a multi-marginal Doob's h-transform to construct a continuous dynamical system conditioned on these irregular observations. Following this, we introduce a variational inference algorithm with a tight evidence lower bound (ELBO), leveraging stochastic optimal control (SOC) theory to approximate the intractable Doob's h-transform and simulate the conditioned dynamics. To improve efficiency and scalability during both training and inference, ACSSM leverages auxiliary variable to flexibly parameterize the latent dynamics and amortized control. Additionally, it incorporates a simulation-free latent dynamics framework and a transformer-based data assimilation scheme, facilitating parallel inference of the latent states and ELBO computation. Through empirical evaluations across a variety of real-world datasets, ACSSM demonstrates superior performance in tasks such as classification, regression, interpolation, and extrapolation, while maintaining computational efficiency.

1749. DS-LLM: Leveraging Dynamical Systems to Enhance Both Training and Inference of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29827 abstract： The training of large language models (LLMs) faces significant computational cost challenges, limiting their scalability toward artificial general intelligence (AGI) and broader adoption. With model sizes doubling approximately every 3.4 months and training costs escalating from 64 million USD for GPT-4 in 2020 to 191 million USD for Gemini Ultra in 2023, the economic burden has become unsustainable. While techniques such as quantization offer incremental improvements, they fail to address the fundamental computational bottleneck. In this work, we introduce DS-LLM, a novel framework that leverages dynamical system (DS)-based machines, which exploit Natural Annealing to rapidly converge to minimal energy states, yielding substantial efficiency gains. Unlike traditional methods, DS-LLM maps LLM components to optimization problems solvable via Hamiltonian configurations and utilizes continuous electric current flow in DS-machines for hardware-native gradient descent during training. We mathematically demonstrate the equivalence between conventional LLMs and DS-LLMs and present a method for transforming a trained LLM into a DS-LLM. Experimental evaluations across multiple model sizes demonstrate orders-of-magnitude improvements in speed and energy efficiency for both training and inference while maintaining consistent accuracy. Additionally, we provide an in-depth analysis of the challenges and potential solutions associated with this emerging computing paradigm, aiming to lay a solid foundation for future research.

1750. Interpreting Emergent Planning in Model-Free Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30433 abstract： We present the first mechanistic evidence that model-free reinforcement learning agents can learn to plan. This is achieved by applying a methodology based on concept-based interpretability to a model-free agent in Sokoban -- a commonly used benchmark for studying planning. Specifically, we demonstrate that DRC, a generic model-free agent introduced by Guez et al. (2019), uses learned concept representations to internally formulate plans that both predict the long-term effects of actions on the environment and influence action selection. Our methodology involves: (1) probing for planning-relevant concepts, (2) investigating plan formation within the agent's representations, and (3) verifying that discovered plans (in the agent's representations) have a causal effect on the agent's behavior through interventions. We also show that the emergence of these plans coincides with the emergence of a planning-like property: the ability to benefit from additional test-time compute. Finally, we perform a qualitative analysis of the planning algorithm learned by the agent and discover a strong resemblance to parallelized bidirectional search. Our findings advance understanding of the internal mechanisms underlying planning behavior in agents, which is important given the recent trend of emergent planning and reasoning capabilities in LLMs through RL.

1751. Robust Weight Initialization for Tanh Neural Networks with Fixed Point

Analysis

链接：https://iclr.cc/virtual/2025/poster/30383 abstract： As a neural network's depth increases, it can improve generalization performance. However, training deep networks is challenging due to gradient and signal propagation issues. To address these challenges, extensive theoretical research and various methods have been introduced. Despite these advances, effective weight initialization methods for tanh neural networks remain insufficiently investigated. This paper presents a novel weight initialization method for neural networks with tanh activation function. Based on an analysis of the fixed points of the function $\tanh(ax)$, the proposed method aims to determine values of a that mitigate activation saturation. A series of experiments on various classification datasets and physics-informed neural networks demonstrates that the proposed method outperforms Xavier initialization methods (with or without normalization) in terms of robustness across different network sizes, data efficiency, and convergence speed. Code is available at https://github.com/1HyunwooLee/Tanh-Init.

1752. Diff-PIC: Revolutionizing Particle-In-Cell Nuclear Fusion Simulation with Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29068 abstract： The rapid development of AI highlights the pressing need for sustainable energy, a critical global challenge for decades. Nuclear fusion, generally seen as a promising solution, has been the focus of intensive research for nearly a century, with investments reaching hundreds of billions of dollars. Recent advancements in Inertial Confinement Fusion (ICF) have drawn significant attention to fusion research, in which Laser-Plasma Interaction (LPI) is critical for ensuring fusion stability and efficiency. However, the complexity of LPI makes analytical approaches impractical, leaving researchers dependent on extremely computationally intensive Particle-in-Cell (PIC) simulations to generate data, posing a significant bottleneck to the advancement of fusion research. In response, this work introduces Diff-PIC, a novel framework that leverages conditional diffusion models as a computationally efficient alternative to PIC simulations for generating high-fidelity scientific LPI data. In this work, physical patterns captured by PIC simulations are distilled into diffusion models associated with two tailored enhancements: (1) To effectively capture the complex relationships between physical parameters and their corresponding outcomes, the parameters are encoded in a physically informed manner. (2) To further enhance efficiency while maintaining physical validity, the rectified flow technique is employed to transform our model into a one-step conditional diffusion model. Experimental results show that Diff-PIC achieves a $\sim$16,200$\times$ speedup compared to traditional PIC on a 100 picosecond simulation, while delivering superior accuracy compared to other data generation approaches.

1753. Effective post-training embedding compression via temperature control in contrastive training

链接：https://iclr.cc/virtual/2025/poster/28082 abstract： Fixed-size learned representations (dense representations, or embeddings) are widely used in many machine learning applications across language, vision or speech modalities. This paper investigates the role of the temperature parameter in contrastive training for text embeddings. We shed light on the impact this parameter has on the intrinsic dimensionality of the embedding spaces obtained, and show that lower intrinsic dimensionality is further correlated with effective compression of embeddings. We still observe a trade-off between absolute performance and effective compression and we propose temperature aggregation methods which reduce embedding size by an order of magnitude with minimal impact on quality.

1754. SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks

链接：https://iclr.cc/virtual/2025/poster/31473 abstract： As the size of large language models continue to scale, so does the computational resources required to run them. Spiking Neural Networks (SNNs) have emerged as an energy-efficient approach to deep learning that leverage sparse and event-driven activations to reduce the computational overhead associated with model inference. While they have become competitive with non-spiking models on many computer vision tasks, SNNs have proven to be more challenging to train. As a result, their performance lags behind modern deep learning, and until now, SNNs have yet to succeed at language generation on large-scale datasets. In this paper, inspired by the Receptance Weighted Key Value (RWKV) language model, we successfully implement `SpikeGPT', a generative language model with binary, event-driven spiking activation units. We train the proposed model on two model variants: 46M and 216M parameters. To the best of our knowledge, SpikeGPT is the largest backpropagation-trained SNN model when released, rendering it suitable for both the generation and comprehension of natural language. We achieve this by modifying the transformer block to replace multi-head self-attention to reduce quadratic computational complexity $\mathcal{O}(T^2)$ to linear complexity $\mathcal{O}(T)$ with increasing sequence length. Input tokens are instead streamed in sequentially to our attention mechanism (as with typical SNNs). Our experiments show that SpikeGPT remains competitive with non-spiking models on tested benchmarks, while maintaining 32.2$\times$ fewer operations when processed on neuromorphic hardware that can leverage sparse, event-driven activations. Our code implementation is available at https://github.com/ridgerchu/SpikeGPT.

1755. In vivo cell-type and brain region classification via multimodal contrastive learning

链接：https://iclr.cc/virtual/2025/poster/31226 abstract： Current electrophysiological approaches can track the activity of many neurons, yet it is usually unknown which cell-types or brain areas are being recorded without further molecular or histological analysis. Developing accurate and scalable algorithms for identifying the cell-type and brain region of recorded neurons is thus crucial for improving our understanding of neural computation. In this work, we develop a multimodal contrastive learning approach for neural data that can be fine-tuned for different downstream tasks, including inference of cell-type and brain location. We utilize multimodal contrastive learning to jointly embed the activity autocorrelations and extracellular waveforms of individual neurons. We demonstrate that our embedding approach, Neuronal Embeddings via MultimOdal Contrastive Learning (NEMO), paired with supervised fine-tuning, achieves state-of-the-art cell-type classification for two opto-tagged datasets and brain region classification for the public International Brain Laboratory Brain-wide Map dataset. Our method represents a promising step towards accurate cell-type and brain region classification from electrophysiological recordings.

1756. Spectral-Refiner: Accurate Fine-Tuning of Spatiotemporal Fourier Neural Operator for Turbulent Flows

链接：https://iclr.cc/virtual/2025/poster/29939 abstract：

1757. LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision

链接：https://iclr.cc/virtual/2025/poster/30575 abstract： Prototypical parts networks combine the power of deep learning with the explainability of case-based reasoning to make accurate, interpretable decisions. They follow the this looks like that reasoning, representing each prototypical part with patches from training images. However, a single image patch comprises multiple visual features, such as color, shape, and texture, making it difficult for users to identify which feature is important to the model.To reduce this ambiguity, we introduce the Lucid Prototypical Parts Network (LucidPPN), a novel prototypical parts network that separates color prototypes from other visual features. Our method employs two reasoning branches: one for non-color visual features, processing grayscale images, and another focusing solely on color information. This separation allows us to clarify whether the model's decisions are based on color, shape, or texture. Additionally, LucidPPN identifies prototypical parts corresponding to semantic parts of classified objects, making comparisons between data classes more intuitive, e.g., when two bird species might differ primarily in belly color.Our experiments demonstrate that the two branches are complementary and together achieve results comparable to baseline methods. More importantly, LucidPPN generates less ambiguous prototypical parts, enhancing user understanding.

1758. DeepLTL: Learning to Efficiently Satisfy Complex LTL Specifications for Multi-Task RL

链接：https://iclr.cc/virtual/2025/poster/30664 abstract： Linear temporal logic (LTL) has recently been adopted as a powerful formalism for specifying complex, temporally extended tasks in multi-task reinforcement learning (RL). However, learning policies that efficiently satisfy arbitrary specifications not observed during training remains a challenging problem. Existing approaches suffer from several shortcomings: they are often only applicable to finite-horizon fragments of LTL, are restricted to suboptimal solutions, and do not adequately handle safety constraints. In this work, we propose a novel learning approach to address these concerns. Our method leverages the structure of Büchi automata, which explicitly represent the semantics of LTL specifications, to learn policies conditioned on sequences of truth assignments that lead to satisfying the desired formulae. Experiments in a variety of discrete and continuous domains demonstrate that our approach is able to zero-shot satisfy a wide range of finite- and infinite-horizon specifications, and outperforms existing methods in terms of both satisfaction probability and efficiency. Code available at: https://deep-ltl.github.io/

1759. FlashRNN: I/O-Aware Optimization of Traditional RNNs on modern hardware

链接：https://iclr.cc/virtual/2025/poster/28546 abstract： While Transformers and other sequence-parallelizable neural network architectures seem like the current state of the art in sequence modeling, they specifically lack state-tracking capabilities. These are important for time-series tasks and logical reasoning. Traditional RNNs like LSTMs and GRUs, as well as modern variants like sLSTM do have these capabilities at the cost of strictly sequential processing. While this is often seen as a strong limitation, we show how fast these networks can get with our hardware-optimization FlashRNN in Triton and CUDA, optimizing kernels to the register level on modern GPUs. We extend traditional RNNs with a parallelization variant that processes multiple RNNs of smaller hidden state in parallel, similar to the head-wise processing in Transformers. To enable flexibility on different GPU variants, we introduce a new optimization framework for hardware-internal cache sizes, memory and compute handling. It models the hardware in a setting using polyhedral-like constraints, including the notion of divisibility. This speeds up the solution process in our ConstrINT library for general integer constraint satisfaction problems (integer CSPs).We show that our kernels can achieve 50x speed-ups over a vanilla PyTorch implementation and allow 40x larger hidden sizes compared to our Triton implementation. We have open-sourced our kernels and the optimization library to boost research in the direction of state-tracking enabled RNNs and sequence modeling here: https://github.com/NX-AI/flashrnn

1760. TokenFormer: Rethinking Transformer Scaling with Tokenized Model Parameters

链接：https://iclr.cc/virtual/2025/poster/28360 abstract： Transformers have become the predominant architecture in foundation models due to their excellent performance across various domains. However, the substantial cost of scaling these models remains a significant concern. This problem arises primarily from their dependence on a fixed number of parameters within linear projections. When architectural modifications (e.g., channel dimensions) are introduced, the entire model typically requires retraining from scratch. As model sizes continue growing, this strategy results in increasingly high computational costs and becomes unsustainable. To overcome this problem, we introduce Tokenformer, a natively scalable architecture that leverages the attention mechanism not only for computations among input tokens but also for interactions between tokens and model parameters, thereby enhancing architectural flexibility. By treating model parameters as tokens, we replace all the linear projections in Transformers with our token-parameter attention layer, where input tokens act as queries and model parameters as keys and values. This reformulation allows for progressive and efficient scaling without necessitating retraining from scratch. Our model scales from 124M to 1.4B parameters by incrementally adding new key-value parameter pairs, achieving performance comparable to Transformers trained from scratch while greatly reducing training costs. Code and models are available at {\color{red}\url{https://github.com/Haiyang-W/TokenFormer.git}}

1761. Vision-LSTM: xLSTM as Generic Vision Backbone

链接：https://iclr.cc/virtual/2025/poster/29584 abstract： Transformers are widely used as generic backbones in computer vision, despite initially introduced for natural language processing. Recently, the Long Short-Term Memory (LSTM) has been extended to a scalable and performant architecture - the xLSTM - which overcomes long-standing LSTM limitations via exponential gating and parallelizable matrix memory structure. In this paper, we introduce Vision-LSTM (ViL), an adaption of the xLSTM building blocks to computer vision. ViL comprises a stack of xLSTM blocks where odd blocks process the sequence of patch tokens from top to bottom while even blocks go from bottom to top.ViL achieves strong performances on classification, transfer learning and segmentation tasks as well as a beneficial pre-training cost-to-performance trade-off. Experiments show that ViL holds promise to be further deployed as new generic backbone for computer vision architectures.

1762. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark

链接：https://iclr.cc/virtual/2025/poster/29528 abstract： The ability to comprehend audio—which includes speech, non-speech sounds, and music—is crucial for AI agents to interact effectively with the world. We present MMAU, a novel benchmark designed to evaluate multimodal audio understanding models on tasks requiring expert-level knowledge and complex reasoning. MMAU comprises 10k carefully curated audio clips paired with human-annotated natural language questions and answers spanning speech, environmental sounds, and music. It includes information extraction and reasoning questions, requiring models to demonstrate 27 distinct skills across unique and challenging tasks. Unlike existing benchmarks, MMAU emphasizes advanced perception and reasoning with domain-specific knowledge, challenging models to tackle tasks akin to those faced by experts. We assess 18 open-source and proprietary (Large) Audio-Language Models, demonstrating the significant challengesposed by MMAU. Notably, even the most advanced Gemini 2.0 Flash achieves only 59.93% accuracy, and the state-of-the-art open-source Qwen2-Audio achieves only 52.50%, highlighting considerable room for improvement. We believe MMAU will drive the audio and multimodal research community to develop more advanced audio understanding models capable of solving complex audio tasks.

1763. Efficiently Parameterized Neural Metriplectic Systems

链接：https://iclr.cc/virtual/2025/poster/27984 abstract： Metriplectic systems are learned from data in a way that scales quadratically in both the size of the state and the rank of the metriplectic operators. In addition to being provably energy-conserving and entropy-stable, the proposed neural metriplectic systems (NMS) approach includes approximation results that demonstrate its ability to accurately learn metriplectic dynamics from data, along with an error estimate that indicates its potential for generalization to unseen timescales when the approximation error is low. Examples are provided to illustrate performance both with full state information available and when entropic variables are unknown, confirming that the NMS approach exhibits superior accuracy and scalability without compromising on model expressivity.

1764. MatExpert: Decomposing Materials Discovery By Mimicking Human Experts

链接：https://iclr.cc/virtual/2025/poster/30631 abstract： Material discovery is a critical research area with profound implications for various industries. In this work, we introduce MatExpert, a novel framework that leverages Large Language Models (LLMs) and contrastive learning to accelerate the discovery and design of new solid-state materials. Inspired by the workflow of human materials design experts, our approach integrates three key stages: retrieval, transition, and generation. First, in the retrieval stage, MatExpert identifies an existing material that closely matches the desired criteria. Second, in the transition stage, MatExpert outlines the necessary modifications to transform this material formulation to meet specific requirements outlined by the initial user query. Third, in the generation state, MatExpert performs detailed computations and

structural generation to create a new material based on the provided information. Our experimental results demonstrate that MatExpert outperforms state-of-the-art methods in material generation tasks, achieving superior performance across various metrics including validity, distribution, and stability. As such, MatExpert represents a meaningful advancement in computational material discovery using langauge-based generative models.

1765. VAE-Var: Variational Autoencoder-Enhanced Variational Methods for Data Assimilation in Meteorology

链接：https://iclr.cc/virtual/2025/poster/27950 abstract： Data assimilation (DA) is an essential statistical technique for generating accurate estimates of a physical system's states by combining prior model predictions with observational data, especially in the realm of weather forecasting. Effectively modeling the prior distribution while adapting to diverse observational sources presents significant challenges for both traditional and neural network-based DA algorithms. This paper introduces VAE-Var, a novel neural network-based data assimilation algorithm aimed at 1) enhancing accuracy by capturing the non-Gaussian characteristics of the conditional background distribution $p(\mathbf{x}|\mathbf{x}_b)$, and 2) efficiently assimilating real-world observational data. VAE-Var utilizes a variational autoencoder to learn the background error distribution, with its decoder creating a variational cost function to optimize the analysis states. The advantages of VAE-Var include: 1) it maintains the framework of traditional variational assimilation, enabling it to accommodate various observation operators, particularly irregular observations; 2) it lessens the dependence on expert knowledge for constructing the background distribution, allowing for improved modeling of non-Gaussian structures; and 3) experimental results indicate that, when applied to the FengWu weather forecasting model, VAE-Var outperforms DiffDA and two traditional algorithms (interpolation and 3DVar) in terms of assimilation accuracy in sparse observational contexts, and is capable of assimilating real-world GDAS prepbufr observations over a year.

1766. Neural Stochastic Differential Equations for Uncertainty-Aware Offline RL

链接：https://iclr.cc/virtual/2025/poster/28729 abstract： Offline model-based reinforcement learning (RL) offers a principled approach to using a learned dynamics model as a simulator to optimize a control policy. Despite the near-optimal performance of existing approaches on benchmarks with high-quality datasets, most struggle on datasets with low state-action space coverage or suboptimal demonstrations.We develop a novel offline model-based RL approach that particularly shines in low-quality data regimes while maintaining competitive performance on high-quality datasets.Neural Stochastic Differential Equations for Uncertainty-aware, Offline RL (NUNO) learns a dynamics model as neural stochastic differential equations (SDE), where its drift term can leverage prior physics knowledge as inductive bias.In parallel, its diffusion term provides distance-aware estimates of model uncertainty by matching the dynamics' underlying stochasticity near the training data regime while providing high but bounded estimates beyond it.To address the so-called model exploitation problem in offline model-based RL, NUNO builds on existing studies by penalizing and adaptively truncating neural SDE's rollouts according to uncertainty estimates.Our empirical results in D4RL and NeoRL MuJoCo benchmarks evidence that NUNO outperforms state-of-the-art methods in low-quality datasets by up to 93% while matching or surpassing their performance by up to 55% in some high-quality counterparts.

1767. Stiefel Flow Matching for Moment-Constrained Structure Elucidation

链接：https://iclr.cc/virtual/2025/poster/30787 abstract： Molecular structure elucidation is a fundamental step in understanding chemical phenomena, with applications in identifying molecules in natural products, lab syntheses, forensic samples, and the interstellar medium.We consider the task of predicting a molecule's all-atom 3D structure given only its molecular formula and moments of inertia, motivated by the ability of rotational spectroscopy to measure these moments.While existing generative models can conditionally sample 3D structures with approximately correct moments, this soft conditioning fails to leverage the many digits of precision afforded by experimental rotational spectroscopy.To address this, we first show that the space of n-atom point clouds with a fixed set of moments of inertia is embedded in the Stiefel manifold $\mathrm{St}(n, 4)$.We then propose Stiefel Flow Matching as a generative model for elucidating 3D structure under exact moment constraints.Additionally, we learn simpler and shorter flows by finding approximate solutions for equivariant optimal transport on the Stiefel manifold.Empirically, enforcing exact moment constraints allows Stiefel Flow Matching to achieve higher success rates and faster sampling than Euclidean diffusion models, even on high-dimensional manifolds corresponding to large molecules in the GEOM dataset.

1768. Robust Root Cause Diagnosis using In-Distribution Interventions

链接：https://iclr.cc/virtual/2025/poster/28541 abstract： Diagnosing the root cause of an anomaly in a complex interconnected system isa pressing problem in today’s cloud services and industrial operations. We propose In-Distribution Interventions (IDI), a novel algorithm that predicts root causeas nodes that meet two criteria: 1) Anomaly: root cause nodes should take onanomalous values; 2) Fix: had the root cause nodes assumed usual values, thetarget node would not have been anomalous. Prior methods of assessing the fixcondition rely on counterfactuals inferred from a Structural Causal Model (SCM)trained on historical data. But since anomalies are rare and fall outside the training distribution, the fitted SCMs yield unreliable counterfactual estimates. IDIovercomes this by relying on interventional estimates obtained by solely probing the fitted SCM at in-distribution inputs. We present a theoretical analysiscomparing and bounding the errors in assessing the fix condition using interventional and counterfactual estimates. We then conduct experiments by systematically varying the SCM’s complexity to demonstrate the

cases where IDI's interventional approach outperforms the counterfactual approach and vice versa.Experiments on both synthetic and PetShop RCD benchmark datasets demonstrate that IDI consistently identifies true root causes more accurately and robustly than nine existing state-of-the-art RCD baselines. Code will be releasedat https://github.com/nlokeshiisc/IDI_release.

1769. CSA: Data-efficient Mapping of Unimodal Features to Multimodal Features

链接：https://iclr.cc/virtual/2025/poster/30894 abstract： Multimodal encoders like CLIP excel in tasks such as zero-shot image classification and cross-modal retrieval. However, they require excessive training data.We propose canonical similarity analysis (CSA), which uses two unimodal encoders to replicate multimodal encoders using limited data.CSA maps unimodal features into a multimodal space, using a new similarity score to retain only the multimodal information.CSA only involves the inference of unimodal encoders and a cubic-complexity matrix decomposition, eliminating the need for extensive GPU-based model training.Experiments show that CSA outperforms CLIP while requiring 50,$000\times$ fewer multimodal data pairs to bridge the modalities given pre-trained unimodal encoders on ImageNet classification and misinformative news caption detection.CSA surpasses the state-of-the-art method to map unimodal features to multimodal features.We also demonstrate the ability of CSA with modalities beyond image and text, paving the way for future modality pairs with limited paired multimodal data but abundant unpaired unimodal data, such as LiDAR and text.

1770. Efficient and Robust Neural Combinatorial Optimization via Wasserstein-Based Coresets

链接：https://iclr.cc/virtual/2025/poster/30367 abstract： Combinatorial optimization (CO) is a fundamental tool in many fields. Many neural combinatorial optimization (NCO) methods have been proposed to solve CO problems.However, existing NCO methods typically require significant computational and storage resources, and face challenges in maintaining robustness to distribution shifts between training and test data.To address these issues, we model CO instances into probability measures, and introduce Wasserstein-based metrics to quantify the difference between CO instances. We then leverage a popular data compression technique, \emph{coreset}, to construct a small-size proxy for the original large dataset.However, the time complexity of constructing a coreset is linearly dependent on the size of the dataset. Consequently, it becomes challenging when datasets are particularly large.Further, we accelerate the coreset construction by adapting it to the merge-and-reduce framework, enabling parallel computing. Additionally, we prove that our coreset is a good representation in theory. {Subsequently}, to speed up the training process for existing NCO methods, we propose an efficient training framework based on the coreset technique. We train the model on a small-size coreset rather than on the full dataset, and thus save substantial computational and storage resources. Inspired by hierarchical Gonzalez's algorithm, our coreset method is designed to capture the diversity of the dataset, which consequently improves robustness to distribution shifts.Finally, experimental results demonstrate that our training framework not only enhances robustness to distribution shifts but also achieves better performance with reduced resource requirements.

1771. Safety-Prioritizing Curricula for Constrained Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28899 abstract： Curriculum learning aims to accelerate reinforcement learning (RL) by generating curricula, i.e., sequences of tasks of increasing difficulty. Although existing curriculum generation approaches provide benefits in sample efficiency, they overlook safety-critical settings where an RL agent must adhere to safety constraints.Thus, these approaches may generate tasks that cause RL agents to violate safety constraints during training and behave suboptimally after. We develop a safe curriculum generation approach (SCG) that aligns the objectives of constrained RL and curriculum learning: improving safety during training and boosting sample efficiency.SCG generates sequences of tasks where the RL agent can be safe and performant by initially generating tasks with minimum safety violations over high-reward ones.We empirically show that compared to the state-of-the-art curriculum learning approaches and their naively modified safe versions, SCG achieves optimal performance and the lowest amount of constraint violations during training.

1772. Semantic Loss Guided Data Efficient Supervised Fine Tuning for Safe Responses in LLMs

链接：https://iclr.cc/virtual/2025/poster/28590 abstract： Large Language Models (LLMs) generating unsafe responses to toxic prompts is a significant issue in their applications. While various efforts aim to address this safety concern, previous approaches often demand substantial human data collection or rely on the less dependable option of using another LLM to generate corrective data. In this paper, we aim to take this problem and overcome limitations of requiring significant high-quality human data. Our method requires only a small set of unsafe responses to toxic prompts, easily obtained from the unsafe LLM itself. By employing a semantic cost combined with a negative Earth Mover Distance (EMD) loss, we guide the LLM away from generating unsafe responses. Additionally, we propose a novel lower bound for EMD loss, enabling more efficient optimization. Our results demonstrate superior performance and data efficiency compared to baselines, and we further examine the nuanced effects of over-alignment and potential degradation of language capabilities when using contrastive data.

1773. Benchmarking LLMs' Judgments with No Gold Standard

链接：https://iclr.cc/virtual/2025/poster/27990 abstract： We introduce the GEM (Generative Estimator for Mutual Information), an evaluation metric for assessing language generation by large language models (LLMs), particularly in generating informative judgments, without the need for a gold standard reference. GEM broadens the scenarios where we can benchmark LLM generation performance-from traditional ones, like machine translation and summarization, where gold standard references are readily available, to subjective tasks without clear gold standards, such as academic peer review.GEM uses a generative model to estimate mutual information between candidate and reference responses, without requiring the reference to be a gold standard. In experiments on two human-annotated datasets, GEM demonstrates competitive correlations with human scores compared to the state-of-the-art GPT-4o Examiner, and outperforms all other baselines. Additionally, GEM is more robust against strategic manipulation, such as rephrasing or elongation, which can artificially inflate scores under a GPT-4o Examiner. We also present GRE-bench (Generating Review Evaluation Benchmark) which evaluates LLMs based on how well they can generate high-quality peer reviews for academic research papers. Because GRE-bench is based upon GEM, it inherits its robustness properties. Additionally, GRE-bench circumvents data contamination problems (or data leakage) by using the continuous influx of new open-access research papers and peer reviews each year. We show GRE-bench results of various popular LLMs on their peer review capabilities using the ICLR2023 dataset.

1774. UIFace: Unleashing Inherent Model Capabilities to Enhance Intra-Class Diversity in Synthetic Face Recognition

链接：https://iclr.cc/virtual/2025/poster/28176 abstract： Face recognition (FR) stands as one of the most crucial applications in computer vision. The accuracy of FR models has significantly improved in recent years due to the availability of large-scale human face datasets. However, directly using these datasets can inevitably lead to privacy and legal problems. Generating synthetic data to train FR models is a feasible solution to circumvent these issues. While existing synthetic-based face recognition methods have made significant progress in generating identity-preserving images, they are severely plagued by context overfitting, resulting in a lack of intra-class diversity of generated images and poor face recognition performance. In this paper, we propose a framework to $\textbf{U}$nleash model $\textbf{I}$nherent capabilities to enhance intra-class diversity for synthetic face recognition, shorted as $\textbf{UIFace}$. Our framework first train a diffusion model that can perform denoising conditioned on either identity contexts or a learnable empty context. The former generates identity-preserving images but lacks variations, while the latter exploits the model's intrinsic ability to synthesize intra-class-diversified images but with random identities. Then we adopt a novel two-stage denoising strategy to fully leverage the strengths of both type of contexts, resulting in images that are diverse as well as identity-preserving. Moreover, an attention injection module is introduced to further augment the intra-class variations by utilizing attention maps from the empty context to guide the denoising process in ID-conditioned generation. Experiments show that our method significantly surpasses previous approaches with even less training data and half the size of synthetic dataset. More surprisingly, the proposed $\textbf{UIFace}$ even achieves comparable performance of FR models trained on real datasets when we increase the number of synthetic identities.

1775. LARP: Tokenizing Videos with a Learned Autoregressive Generative Prior

链接：https://iclr.cc/virtual/2025/poster/29340 abstract： We present LARP, a novel video tokenizer designed to overcome limitations in current video tokenization methods for autoregressive (AR) generative models. Unlike traditional patchwise tokenizers that directly encode local visual patches into discrete tokens, LARP introduces a holistic tokenization scheme that gathers information from the visual content using a set of learned holistic queries. This design allows LARP to capture more global and semantic representations, rather than being limited to local patch-level information. Furthermore, it offers flexibility by supporting an arbitrary number of discrete tokens, enabling adaptive and efficient tokenization based on the specific requirements of the task. To align the discrete token space with downstream AR generation tasks, LARP integrates a lightweight AR transformer as a training-time prior model that predicts the next token on its discrete latent space. By incorporating the prior model during training, LARP learns a latent space that is not only optimized for video reconstruction but is also structured in a way that is more conducive to autoregressive generation. Moreover, this process defines a sequential order for the discrete tokens, progressively pushing them toward an optimal configuration during training, ensuring smoother and more accurate AR generation at inference time. Comprehensive experiments demonstrate LARPs strong performance, achieving state-of-the-art FVD on the UCF101 class-conditional video generation benchmark. LARP enhances the compatibility of AR models with videos and opens up the potential to build unified high-fidelity multimodal large language models (MLLMs). Project page: https://hywang66.github.io/larp/

1776. Transformers Provably Learn Two-Mixture of Linear Classification via Gradient Flow

链接：https://iclr.cc/virtual/2025/poster/30607 abstract： Understanding how transformers learn and utilize hidden connections between tokens is crucial to understand the behavior of large language models.To understand this mechanism, we consider the task of two-mixture of linear classification which possesses a hidden correspondence structure among tokens, and study the training dynamics of a symmetric two-headed transformer with ReLU neurons.Motivated by the stage-wise learning phenomenon in our experiments, we design and theoretically analyze a three-stage training algorithm, which can effectively characterize the actual gradient descent dynamics when we simultaneously train the neuron weights and the softmax attention.The first stage is a neuron learning stage, where the neurons align with the underlying signals. The second stage is a attention feature learning stage, where we analyze the feature learning process of how the attention learns to utilize the relationship between the tokens to

solve certain hard samples.In the meantime, the attention features evolve from a nearly non-separable state (at the initialization) to a well-separated state.The third stage is a convergence stage, where the population loss is driven towards zero.The key technique in our analysis of softmax attention is to identify a critical sub-system inside a large dynamical system and bound the growth of the non-linear sub-system by a linear system. Finally, we discuss the setting with more than two mixtures. We empirically show the difficulty of generalizing our analysis of the gradient flow dynamics to the case even when the number of mixtures equals three, although the transformer can still successfully learn such distribution. On the other hand, we show by construction that there exists a transformer that can solve mixture of linear classification given any arbitrary number of mixtures.

1777. Forewarned is Forearmed: Harnessing LLMs for Data Synthesis via Failure-induced Exploration

链接：https://iclr.cc/virtual/2025/poster/27706 abstract： Large language models (LLMs) have significantly benefited from training on diverse, high-quality task-specific data, leading to impressive performance across a range of downstream applications. Current methods often rely on human-annotated data or predefined task templates to direct powerful LLMs in synthesizing task-relevant data for effective model training. However, this dependence on manually designed components may constrain the scope of generated data, potentially overlooking critical edge cases or novel scenarios that could challenge the model. In this paper, we present a novel approach, ReverseGen, designed to automatically generate effective training samples that expose the weaknesses of LLMs. Specifically, we introduce a dedicated proposer trained to produce queries that lead target models to generate unsatisfactory responses. These failure-inducing queries are then used to construct training data, helping to address the models' shortcomings and improve overall performance. Our approach is flexible and can be applied to models of various scales (3B, 7B, and 8B). We evaluate ReverseGen on three key applications—safety, honesty, and math—demonstrating that our generated data is both highly effective and diverse. Models fine-tuned with ReverseGen-generated data consistently outperform those trained on human-annotated or general model-generated data, offering a new perspective on data synthesis for task-specific LLM enhancement.

1778. CheapNet: Cross-attention on Hierarchical representations for Efficient protein-ligand binding Affinity Prediction

链接：https://iclr.cc/virtual/2025/poster/30656 abstract： Accurately predicting protein-ligand binding affinity is a critical challenge in drug discovery, crucial for understanding drug efficacy. While existing models typically rely on atom-level interactions, they often fail to capture the complex, higher-order interactions, resulting in noise and computational inefficiency. Transitioning to modeling these interactions at the cluster level is challenging because it is difficult to determine which atoms form meaningful clusters that drive the protein-ligand interactions. To address this, we propose CheapNet, a novel interaction-based model that integrates atom-level representations with hierarchical cluster-level interactions through a cross-attention mechanism. By employing differentiable pooling of atom-level embeddings, CheapNet efficiently captures essential higher-order molecular representations crucial for accurate binding predictions. Extensive evaluations demonstrate that CheapNet not only achieves state-of-the-art performance across multiple binding affinity prediction tasks but also maintains prediction accuracy with reasonable computational efficiency. The code of CheapNet is available at https://github.com/hyukjunlim/CheapNet.

1779. Provable Uncertainty Decomposition via Higher-Order Calibration

链接：https://iclr.cc/virtual/2025/poster/29551 abstract： We give a principled method for decomposing the predictive uncertainty of a model into aleatoric and epistemic components with explicit semantics relating them to the real-world data distribution. While many works in the literature have proposed such decompositions, they lack the type of formal guarantees we provide. Our method is based on the new notion of higher-order calibration, which generalizes ordinary calibration to the setting of higher-order predictors that predict *mixtures* over label distributions at every point. We show how to measure as well as achieve higher-order calibration using access to k-snapshots, namely examples where each point has k independent conditional labels. Under higher-order calibration, the estimated aleatoric uncertainty at a point is guaranteed to match the real-world aleatoric uncertainty averaged over all points where the prediction is made. To our knowledge, this is the first formal guarantee of this type that places no assumptions whatsoever on the real-world data distribution. Importantly, higher-order calibration is also applicable to existing higher-order predictors such as Bayesian and ensemble models and provides a natural evaluation metric for such models. We demonstrate through experiments that our method produces meaningful uncertainty decompositions in tasks such as image classification.

1780. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models

链接：https://iclr.cc/virtual/2025/poster/30192 abstract： We introduce methods for discovering and applying sparse feature circuits. These are causally implicated subnetworks of human-interpretable features for explaining language model behaviors. Circuits identified in prior work consist of polysemantic and difficult-to-interpret units like attention heads or neurons, rendering them unsuitable for many downstream applications. In contrast, sparse feature circuits enable detailed understanding of unanticipated mechanisms in neural networks. Because they are based on fine-grained units, sparse feature circuits are useful for downstream tasks: We introduce SHIFT, where we improve the generalization of a classifier by ablating features that a human judges to be task-irrelevant. Finally, we demonstrate an entirely unsupervised and scalable interpretability pipeline by

discovering thousands of sparse feature circuits for automatically discovered model behaviors.

1781. SePer: Measure Retrieval Utility Through The Lens Of Semantic Perplexity Reduction

链接：https://iclr.cc/virtual/2025/poster/28663 abstract： Large Language Models (LLMs) have demonstrated improved generation performance by incorporating externally retrieved knowledge, a process known as retrieval-augmented generation (RAG). Despite the potential of this approach, existing studies evaluate RAG effectiveness by 1) assessing retrieval and generation components jointly, which obscures retrieval's distinct contribution, or 2) examining retrievers using traditional metrics such as NDCG, which creates a gap in understanding retrieval's true utility in the overall generation process. To address the above limitations, in this work, we introduce an automatic evaluation method that measures retrieval quality through the lens of information gain within the RAG framework. Specifically, we propose Semantic Perplexity (SePer), a metric that captures the LLM's internal belief about the correctness of the retrieved information. We quantify the utility of retrieval by the extent to which it reduces semantic perplexity post-retrieval. Extensive experiments demonstrate that SePer not only aligns closely with human preferences but also offers a more precise and efficient evaluation of retrieval utility across diverse RAG scenarios.

1782. A Generic Framework for Conformal Fairness

链接：https://iclr.cc/virtual/2025/poster/27770 abstract： Conformal Prediction (CP) is a popular method for uncertainty quantification with machine learning models. While conformal prediction provides probabilistic guarantees regarding the coverage of the true label, these guarantees are agnostic to the presence of sensitive attributes within the dataset. In this work, we formalize \textit{Conformal Fairness}, a notion of fairness using conformal predictors, and provide a theoretically well-founded algorithm and associated framework to control for the gaps in coverage between different sensitive groups. Our framework leverages the exchangeability assumption (implicit to CP) rather than the typical IID assumption, allowing us to apply the notion of Conformal Fairness to data types and tasks that are not IID, such as graph data. Experiments were conducted on graph and tabular datasets to demonstrate that the algorithm can control fairness-related gaps in addition to coverage aligned with theoretical expectations.

1783. Representative Guidance: Diffusion Model Sampling with Coherence

链接：https://iclr.cc/virtual/2025/poster/28815 abstract： The diffusion sampling process faces a persistent challenge stemming from its incoherence, attributable to varying noise directions across different timesteps.Our Representative Guidance (RepG) offers a new perspective to address this issue by reformulating the sampling process with a coherent direction toward a representative target.From this perspective, classic classifier guidance reveals its drawback in lacking meaningful representative information, as the features it relies on are optimized for discrimination and tend to highlight only a narrow set of class-specific cues. This focus often sacrifices diversity and increases the risk of adversarial generation.In contrast, we leverage self-supervised representations as the coherent target and treat sampling as a downstream task—one that focuses on refining image details and correcting generation errors, rather than settling for oversimplified outputs.Our Representative Guidance achieves superior performance and demonstrates the potential of pre-trained self-supervised models in guiding diffusion sampling. Our findings show that RepG not only significantly improves vanilla diffusion sampling, but also surpasses state-of-the-art benchmarks when combined with classifier-free guidance.

1784. Spread Preference Annotation: Direct Preference Judgment for Efficient LLM Alignment

链接：https://iclr.cc/virtual/2025/poster/30572 abstract： Aligning large language models (LLMs) with human preferences becomes a key component to obtaining state-of-the-art performance, but it yields a huge cost to construct a large human-annotated preference dataset. To tackle this problem, we propose a new framework, Spread Preference Annotation with direct preference judgment (SPA), that boosts the alignment of LLMs using only a very small amount of human-annotated preference data.Our key idea is leveraging the human prior knowledge within the small (seed) data and progressively improving the alignment of LLM, by iteratively generating the responses and learning from them with the self-annotated preference data.To be specific, we propose to derive the preference label from the logits of LLM to explicitly extract the model's inherent preference. Compared to the previous approaches using external reward models or implicit in-context learning, we observe that the proposed approach is significantly more effective.In addition, we introduce a noise-aware preference learning algorithm to mitigate the risk of low quality within generated preference data.Our experimental results demonstrate that the proposed framework significantly boosts the alignment of LLMs.For example, we achieve superior alignment performance on AlpacaEval 2.0 with only 3.3% of the ground-truth preference labels in the Ultrafeedback data compared to the cases using the entire data or state-of-the-art baselines.

1785. Swiss Army Knife: Synergizing Biases in Knowledge from Vision Foundation Models for Multi-Task Learning

链接：https://iclr.cc/virtual/2025/poster/28912 abstract： Vision Foundation Models (VFMs) have demonstrated outstanding performance on numerous downstream tasks. However, due to their inherent representation biases originating from different

training paradigms, VFMs exhibit advantages and disadvantages across distinct vision tasks. Although amalgamating the strengths of multiple VFMs for downstream tasks is an intuitive strategy, effectively exploiting these biases remains a significant challenge. In this paper, we propose a novel and versatile "Swiss Army Knife" (SAK) solution, which adaptively distills knowledge from a committee of VFMs to enhance multi-task learning. Unlike existing methods that use a single backbone for knowledge transfer, our approach preserves the unique representation bias of each teacher by collaborating the lightweight Teacher-Specific Adapter Path modules with the Teacher-Agnostic Stem. Through dynamic selection and combination of representations with Mixture-of-Representations Routers, our SAK is capable of synergizing the complementary strengths of multiple VFMs. Extensive experiments show that our SAK remarkably outperforms prior state of the arts in multi-task learning by 10% on the NYUD-v2 benchmark, while also providing a flexible and robust framework that can readily accommodate more advanced model designs.

1786. Optimizing importance weighting in the presence of sub-population shifts

链接：https://iclr.cc/virtual/2025/poster/28656 abstract： A distribution shift between the training and test data can severely harm performance of machine learning models. Importance weighting addresses this issue by assigning different weights to data points during training. We argue that existing heuristics for determining the weights are suboptimal, as they neglect the increase of the variance of the estimated model due to the limited sample size of the training data. We interpret the optimal weights in terms of a bias-variance trade-off, and propose a bi-level optimization procedure in which the weights and model parameters are optimized simultaneously. We apply this framework to existing importance weighting techniques for last-layer retraining of deep neural networks in the presence of sub-population shifts and show empirically that optimizing weights significantly improves generalization performance.

1787. Synthesizing Realistic fMRI: A Physiological Dynamics-Driven Hierarchical Diffusion Model for Efficient fMRI Acquisition

链接：https://iclr.cc/virtual/2025/poster/27664 abstract： Functional magnetic resonance imaging (fMRI) is essential for mapping brain activity but faces challenges like lengthy acquisition time and sensitivity to patient movement, limiting its clinical and machine learning applications. While generative models such as diffusion models can synthesize fMRI signals to alleviate these issues, they often underperform due to neglecting the brain's complex structural and dynamic properties.To address these limitations, we propose the Physiological Dynamics-Driven Hierarchical Diffusion Model, a novel framework integrating two key brain physiological properties into the diffusion process: brain hierarchical regional interactions and multifractal dynamics. To model complex interactions among brain regions, we construct hypergraphs based on the prior knowledge of brain functional parcellation reflected by resting-state functional connectivity (rsFC). This enables the aggregation of fMRI signals across multiple scales and generates hierarchical signals. Additionally, by incorporating the prediction of two key dynamics properties of fMRI—the multifractal spectrum and generalized Hurst exponent—our framework effectively guides the diffusion process, ensuring the preservation of the scale-invariant characteristics inherent in real fMRI data.Our framework employs progressive diffusion generation, with signals representing broader brain region information conditioning those that capture localized details, and unifies multiple inputs during denoising for balanced integration.Experiments demonstrate that our model generates physiologically realistic fMRI signals, potentially reducing acquisition time and enhancing data quality, benefiting clinical diagnostics and machine learning in neuroscience.

1788. Logical Consistency of Large Language Models in Fact-Checking

链接：https://iclr.cc/virtual/2025/poster/29583 abstract： In recent years, large language models (LLMs) have demonstrated significant success in performing varied natural language tasks such as language translation, question-answering, summarizing, fact-checking, etc. Despite LLMs' impressive ability to generate human-like texts, LLMs are infamous for their inconsistent responses – a meaning-preserving change in the input query results in an inconsistent response and attributes to vulnerabilities of LLMs such as hallucination. Consequently, existing research focuses on simple paraphrasing-based consistency assessment of LLMs, and ignores complex queries that necessitate an even better understanding of logical reasoning by an LLM. Our work therefore addresses the logical inconsistency of LLMs under complex logical queries with primitive logical operators, e.g., negation, conjunction, and disjunction. As a test bed, we consider retrieval-augmented LLMs on a fact-checking task involving propositional logic queries from knowledge graphs (KGs). Our contributions are three-fold. Benchmark: We introduce three logical fact-checking datasets over KGs for community development towards logically consistent LLMs. Assessment: We propose consistency measures of LLMs on propositional logic queries and demonstrate that existing LLMs lack logical consistency, especially on complex queries. Improvement: We employ supervised fine-tuning to improve the logical consistency of LLMs on the complex fact-checking task with KG contexts. We have made our source code and benchmarks available.

1789. ProtComposer: Compositional Protein Structure Generation with 3D Ellipsoids

链接：https://iclr.cc/virtual/2025/poster/31248 abstract： We develop ProtComposer to generate protein structures conditioned on spatial protein layouts that are specified via a set of 3D ellipsoids capturing substructure shapes and semantics. At inference time, we condition on ellipsoids that are hand-constructed, extracted from existing proteins, or from a statistical model, with each

option unlocking new capabilities. Hand-specifying ellipsoids enables users to control the location, size, orientation, secondary structure, and approximate shape of protein substructures. Conditioning on ellipsoids of existing proteins enables redesigning their substructure's connectivity or editing substructure properties. By conditioning on novel and diverse ellipsoid layouts from a simple statistical model, we improve protein generation with expanded Pareto frontiers between designability, novelty, and diversity. Further, this enables sampling designable proteins with a helix-fraction that matches PDB proteins, unlike existing generative models that commonly oversample conceptually simple helix bundles. Code is available at https://github.com/NVlabs/protcomposer.

1790. TAU-106K: A New Dataset for Comprehensive Understanding of Traffic Accident

链接：https://iclr.cc/virtual/2025/poster/32101 abstract： Multimodal Large Language Models (MLLMs) have demonstrated impressive performance in general visual understanding tasks. However, their potential for high-level, fine-grained comprehension, such as anomaly understanding, remains unexplored. Focusing on traffic accidents, a critical and practical scenario within anomaly understanding, we investigate the advanced capabilities of MLLMs and propose TABot, a multimodal MLLM specialized for accident-related tasks. To facilitate this, we first construct TAU-106K, a large-scale multimodal dataset containing 106K traffic accident videos and images collected from academic benchmarks and public platforms. The dataset is meticulously annotated through a video-to-image annotation pipeline to ensure comprehensive and high-quality labels. Building upon TAU-106K, we train TABot using a two-step approach designed to integrate multi-granularity tasks, including accident recognition, spatial-temporal grounding, and an auxiliary description task to enhance the model's understanding of accident elements. Extensive experiments demonstrate TABot's superior performance in traffic accident understanding, highlighting not only its capabilities in high-level anomaly comprehension but also the robustness of the TAU-106K benchmark. Our code and data will be available at https://github.com/cool-xuan/TABot.

1791. Conditional Diffusion Models are Minimax-Optimal and Manifold-Adaptive for Conditional Distribution Estimation

链接：https://iclr.cc/virtual/2025/poster/29860 abstract： We consider a class of conditional forward-backward diffusion models for conditional generative modeling, that is, generating new data given a covariate (or control variable). To formally study the theoretical properties of these conditional generative models, we adopt a statistical framework of distribution regression to characterize the large sample properties of the conditional distribution estimators induced by these conditional forward-backward diffusion models. Here, the conditional distribution of data is assumed to smoothly change over the covariate. In particular, our derived convergence rate is minimax-optimal under the total variation metric within the regimes covered by the existing literature. Additionally, we extend our theory by allowing both the data and the covariate variable to potentially admit a low-dimensional manifold structure. In this scenario, we demonstrate that the conditional forward-backward diffusion model can adapt to both manifold structures, meaning that the derived estimation error bound (under the Wasserstein metric) depends only on the intrinsic dimensionalities of the data and the covariate.

1792. CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs

链接：https://iclr.cc/virtual/2025/poster/30804 abstract： Multimodal Large Language Models (MLLMs) still struggle with hallucinations despite their impressive capabilities. Recent studies have attempted to mitigate this by applying Direct Preference Optimization (DPO) to multimodal scenarios using preference pairs from text-based responses. However, our analysis of representation distributions reveals that multimodal DPO struggles to align image and text representations and to distinguish between hallucinated and non-hallucinated descriptions. To address these challenges,In this work, we propose a Cross-modal Hierarchical Direct Preference Optimization (CHiP) to address these limitations.We introduce a visual preference optimization module within the DPO framework, enabling MLLMs to learn from both textual and visual preferences simultaneously. Furthermore, we propose a hierarchical textual preference optimization module that allows the model to capture preferences at multiple granular levels, including response, segment, and token levels. We evaluate CHiP through both quantitative and qualitative analyses, with results across multiple benchmarks demonstrating its effectiveness in reducing hallucinations. On the Object HalBench dataset, CHiP outperforms DPO in hallucination reduction, achieving improvements of 52.7% and 55.5% relative points based on the base model Muffin and LLaVA models, respectively. We make all our datasets and code publicly available.

1793. HARDMath: A Benchmark Dataset for Challenging Problems in Applied Mathematics

链接：https://iclr.cc/virtual/2025/poster/28419 abstract： Advanced applied mathematics problems are underrepresented in existing Large Language Model (LLM) benchmark datasets. To address this, we introduce $\textbf{HARDMath}$, a dataset inspired by a graduate course on asymptotic methods, featuring challenging applied mathematics problems that require analytical approximation techniques. These problems demand a combination of mathematical reasoning, computational tools, and subjective judgment, making them difficult for LLMs. Our framework auto-generates a large number of problems with solutions validated against numerical ground truths. We evaluate both open- and closed-source LLMs on $\textbf{HARDMath-

mini}$, a sub-sampled test set of 366 problems, as well as on 40 word problems formulated in applied science contexts. Even leading closed-source models like GPT-4 achieve only 43.8% overall accuracy with few-shot Chain-of-Thought prompting, and all models demonstrate significantly lower performance compared to results on existing mathematics benchmark datasets. We additionally conduct a detailed error analysis to gain insights into the failure cases of LLMs. These results demonstrate the limitations of current LLM performance on advanced graduate-level applied math problems and underscore the importance of datasets like $\textbf{HARDMath}$ to advance mathematical abilities of LLMs.

1794. Multimodal Lego: Model Merging and Fine-Tuning Across Topologies and Modalities in Biomedicine

链接：https://iclr.cc/virtual/2025/poster/28308 abstract： Learning holistic computational representations in physical, chemical or biological systems requires the ability to process information from different distributions and modalities within the same model. Thus, the demand for multimodal machine learning models has sharply risen for modalities that go beyond vision and language, such as sequences, graphs, time series, or tabular data. While there are many available multimodal fusion and alignment approaches, most of them require end-to-end training, scale quadratically with the number of modalities, cannot handle cases of high modality imbalance in the training set, or are highly topology-specific, making them too restrictive for many biomedical learning tasks. This paper presents Multimodal Lego (MM-Lego), a general-purpose fusion framework to turn any set of encoders into a competitive multimodal model with no or minimal fine-tuning. We achieve this by introducing a wrapper for any unimodal encoder that enforces shape consistency between modality representations. It harmonises these representations by learning features in the frequency domain to enable model merging with little signal interference. We show that MM-Lego 1) can be used as a model merging method which achieves competitive performance with end-to-end fusion models without any fine-tuning, 2) can operate on any unimodal encoder, and 3) is a model fusion method that, with minimal fine-tuning, surpasses all benchmarks in five out of seven datasets.

1795. Intent3D: 3D Object Detection in RGB-D Scans Based on Human Intention

链接：https://iclr.cc/virtual/2025/poster/30957 abstract： In real-life scenarios, humans seek out objects in the 3D world to fulfill their daily needs or intentions. This inspires us to introduce 3D intention grounding, a new task in 3D object detection employing RGB-D, based on human intention, such as "I want something to support my back." Closely related, 3D visual grounding focuses on understanding human reference. To achieve detection based on human intention, it relies on humans to observe the scene, reason out the target that aligns with their intention ("pillow" in this case), and finally provide a reference to the AI system, such as "A pillow on the couch". Instead, 3D intention grounding challenges AI agents to automatically observe, reason and detect the desired target solely based on human intention. To tackle this challenge, we introduce the new Intent3D dataset, consisting of 44,990 intention texts associated with 209 fine-grained classes from 1,042 scenes of the ScanNet dataset. We also establish several baselines based on different language-based 3D object detection models on our benchmark. Finally, we propose IntentNet, our unique approach, designed to tackle this intention-based detection problem. It focuses on three key aspects: intention understanding, reasoning to identify object candidates, and cascaded adaptive learning that leverages the intrinsic priority logic of different losses for multiple objective optimization.

1796. Elliptic Loss Regularization

链接：https://iclr.cc/virtual/2025/poster/29240 abstract： Regularizing neural networks is important for anticipating model behavior in regions of the data space that are not well represented. In this work, we propose a regularization technique for enforcing a level of smoothness in the mapping between the input space and the loss. We specify the level of regularity by requiring that the loss of the network satisfies an elliptic operator over the data domain. To do this, we modify the usual empirical risk minimization objective such that we instead minimize a new objective that satisfies an elliptic operator over points within the domain. This allows us to use existing theory on elliptic operators to anticipate the behavior of the error for points outside the training set. We propose a tractable computational method that approximates the behavior of the elliptic operator while being computationally efficient. Finally, we analyze the properties of the proposed regularization to understand the performance on common problems of distribution shift and group imbalance. Numerical experiments empirically confirm the promise of the proposed regularization technique.

1797. Tuning-Free Bilevel Optimization: New Algorithms and Convergence Analysis

链接：https://iclr.cc/virtual/2025/poster/30652 abstract： Bilevel optimization has recently attracted considerable attention due to its abundant applications in machine learning problems. However, existing methods rely on prior knowledge of problem parameters to determine stepsizes, resulting in significant effort in tuning stepsizes when these parameters are unknown. In this paper, we propose two novel tuning-free algorithms, D-TFBO and S-TFBO. D-TFBO employs a double-loop structure with stepsizes adaptively adjusted by the "inverse of cumulative gradient norms" strategy. S-TFBO features a simpler fully single-loop structure that updates three variables simultaneously with a theory-motivated joint design of adaptive stepsizes for all variables. We provide a comprehensive convergence analysis for both algorithms and show that D-TFBO and S-TFBO respectively require $\mathcal{O}(\frac{1}{\epsilon})$ and $\mathcal{O}(\frac{1}{\epsilon}\log^4(\frac{1}{\epsilon}))$ iterations to find an

ϵ-accurate stationary point, (nearly) matching their well-tuned counterparts using the information of problem parameters. Experiments on various problems show that our methods achieve performance comparable to existing well-tuned approaches, while being more robust to the selection of initial stepsizes. To the best of our knowledge, our methods are the first to completely eliminate the need for stepsize tuning, while achieving theoretical guarantees.

1798. How to visualize training dynamics in neural networks

链接：https://iclr.cc/virtual/2025/poster/31343 abstract： Deep learning practitioners typically rely on training and validation loss curves to understand neural network training dynamics. This blog post demonstrates how classical data analysis tools like PCA and hidden Markov models can reveal how neural networks learn different data subsets and identify distinct training phases. We show that traditional statistical methods remain valuable for understanding the training dynamics of modern deep learning systems.

1799. Progressive Compression with Universally Quantized Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30486 abstract： Diffusion probabilistic models have achieved mainstream success in many generative modeling tasks, from image generation to inverse problem solving. A distinct feature of these models is that they correspond to deep hierarchical latent variable models optimizing a variational evidence lower bound (ELBO) on the data likelihood. Drawing on a basic connection between likelihood modeling and compression, we explore the potential of diffusion models for progressive coding, resulting in a sequence of bits that can be incrementally transmitted and decoded with progressively improving reconstruction quality. Unlike prior work based on Gaussian diffusion or conditional diffusion models, we propose a new form of diffusion model with uniform noise in the forward process, whose negative ELBO corresponds to the end-to-end compression cost using universal quantization. We obtain promising first results on image compression, achieving competitive rate-distortion-realism results on a wide range of bit-rates with a single model, bringing neural codecs a step closer to practical deployment. Our code can be found at https://github.com/mandt-lab/uqdm.

1800. AstroCompress: A benchmark dataset for multi-purpose compression of astronomical data

链接：https://iclr.cc/virtual/2025/poster/28585 abstract： The site conditions that make astronomical observatories in space and on the ground so desirable---cold and dark---demand a physical remoteness that leads to limited data transmission capabilities. Such transmission limitations directly bottleneck the amount of data acquired and in an era of costly modern observatories, any improvements in lossless data compression has the potential scale to billions of dollars worth of additional science that can be accomplished on the same instrument. Traditional lossless methods for compressing astrophysical data are manually designed. Neural data compression, on the other hand, holds the promise of learning compression algorithms end-to-end from data and outperforming classical techniques by leveraging the unique spatial, temporal, and wavelength structures of astronomical images. This paper introduces AstroCompress: a neural compression challenge for astrophysics data, featuring four new datasets (and one legacy dataset) with 16-bit unsigned integer imaging data in various modes: space-based, ground-based, multi-wavelength, and time-series imaging. We provide code to easily access the data and benchmark seven lossless compression methods (three neural and four non-neural, including all practical state-of-the-art algorithms).Our results on lossless compression indicate that lossless neural compression techniques can enhance data collection at observatories, and provide guidance on the adoption of neural compression in scientific applications. Though the scope of this paper is restricted to lossless compression, we also comment on the potential exploration of lossy compression methods in future studies.