1201. Federated Few-Shot Class-Incremental Learning

链接：https://iclr.cc/virtual/2025/poster/29204 abstract： This study proposes a challenging yet practical Federated Few-Shot Class-Incremental Learning (FFSCIL) problem, where clients only hold very few samples for new classes. We develop a novel Unified Optimized Prototype Prompt (UOPP) model to simultaneously handle catastrophic forgetting, over-fitting, and prototype bias in FFSCIL. UOPP utilizes task-wise prompt learning to mitigate task interference and over-fitting, unified static-dynamic prototypes to achieve a stability-plasticity balance, and adaptive dual heads for enhanced inferences. Dynamic prototypes represent new classes in the current few-shot task and are rectified to deal with prototype bias. Our comprehensive experimental results show that UOPP significantly outperforms state-of-the-art (SOTA) methods on three datasets with improvements up to 76% on average accuracy and 90% on harmonic mean accuracy respectively. Our extensive analysis shows UOPP robustness in various numbers of local clients and global rounds, low communication costs, and moderate running time. The source code of UOPP is publicly available at https://github.com/anwarmaxsum/FFSCIL.

1202. Contractive Dynamical Imitation Policies for Efficient Out-of-Sample Recovery

链接：https://iclr.cc/virtual/2025/poster/28526 abstract： Imitation learning is a data-driven approach to learning policies from expert behavior, but it is prone to unreliable outcomes in out-of-sample (OOS) regions. While previous research relying on stable dynamical systems guarantees convergence to a desired state, it often overlooks transient behavior. We propose a framework for learning policies modeled by contractive dynamical systems, ensuring that all policy rollouts converge regardless of perturbations, and in turn, enable efficient OOS recovery. By leveraging recurrent equilibrium networks and coupling layers, the policy structure guarantees contractivity for any parameter choice, which facilitates unconstrained optimization. We also provide theoretical upper bounds for worst-case and expected loss to rigorously establish the reliability of our method in deployment. Empirically, we demonstrate substantial OOS performance improvements for simulated robotic manipulation and navigation tasks. See sites.google.com/view/contractive-dynamical-policies for our codebase and highlight of the results.

1203. Inference Scaling for Long-Context Retrieval Augmented Generation

链接：https://iclr.cc/virtual/2025/poster/30339 abstract： The scaling of inference computation has unlocked the potential of long-context large language models (LLMs) across diverse settings. For knowledge-intensive tasks, the increased compute is often allocated to incorporate more external knowledge. However, without effectively utilizing such knowledge, solely expanding context does not always enhance performance. In this work, we investigate inference scaling for retrieval augmented generation (RAG), exploring the combination of multiple strategies beyond simply increasing the quantity of knowledge, including in-context learning and iterative prompting. These strategies provide additional flexibility to scale test-time computation (e.g., by increasing retrieved documents or generation steps), thereby enhancing LLMs' ability to effectively acquire and utilize contextual information. We address two key questions: (1) How does RAG performance benefit from the scaling of inference computation when optimally configured? (2) Can we predict the optimal test-time compute allocation for a given budget by modeling the relationship between RAG performance and inference parameters? Our observations reveal that increasing inference computation leads to nearly linear gains in RAG performance when optimally allocated, a relationship we describe as the inference scaling laws for RAG. Building on this, we further develop the computation allocation model to estimate RAG performance across different inference configurations. The model predicts optimal inference parameters under various computation constraints, which align closely with the experimental results. By applying these optimal configurations, we demonstrate that scaling inference compute on long-context LLMs achieves up to 58.9% gains on benchmark datasets compared to standard RAG.

1204. Vision and Language Synergy for Rehearsal Free Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30681 abstract： The prompt-based approach has demonstrated its success for continual learning problems. However, it still suffers from catastrophic forgetting due to inter-task vector similarity and unfitted new components of previously learned tasks. On the other hand, the language-guided approach falls short of its full potential due to minimum utilized knowledge and participation in the prompt tuning process. To correct this problem, we propose a novel prompt-based structure and algorithm that incorporate 4 key concepts (1) language as input for prompt generation (2) task-wise generators (3) limiting matching descriptors search space via soft task-id prediction (4) generated prompt as auxiliary data. Our experimental analysis shows the superiority of our method to existing SOTAs in CIFAR100, ImageNet-R, and CUB datasets with significant margins i.e. up to 30% final average accuracy, 24% cumulative average accuracy, 8% final forgetting measure, and 7% cumulative forgetting measure. Our historical analysis confirms our method successfully maintains the stability-plasticity trade-off in every task. Our robustness analysis shows the proposed method consistently achieves high performances in various prompt lengths, layer depths, and number of generators per task compared to the SOTAs. We provide a comprehensive theoretical analysis, and complete numerical results in appendix sections. The method code is available in https://github.com/anwarmaxsum/LEAPGEN for further study.

1205. Large Language Models Assume People are More Rational than We Really are

链接：https://iclr.cc/virtual/2025/poster/29001 abstract： In order for AI systems to communicate effectively with people, they must understand how we make decisions. However, people's decisions are not always rational, so the implicit internal models of human decision-making in Large Language Models (LLMs) must account for this. Previous empirical evidence seems to suggest that these implicit models are accurate --- LLMs offer believable proxies of human behavior, acting how we expect humans would in everyday interactions. However, by comparing LLM behavior and predictions to a large dataset of human decisions, we find that this is actually not the case: when both simulating and predicting people's choices, a suite of cutting-edge LLMs (GPT-4o \& 4-Turbo, Llama-3-8B \& 70B, Claude 3 Opus) assume that people are more rational than we really are. Specifically, these models deviate from human behavior and align more closely with a classic model of rational choice --- expected value theory. Interestingly, people also tend to assume that other people are rational when interpreting their behavior. As a consequence, when we compare the inferences that LLMs and people draw from the decisions of others using another psychological dataset, we find that these inferences are highly correlated. Thus, the implicit decision-making models of LLMs appear to be aligned with the human expectation that other people will act rationally, rather than with how people actually act.

1206. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge

链接：https://iclr.cc/virtual/2025/poster/30419 abstract： The effectiveness of automatic evaluation of generative models is typically measured by comparing the labels generated via automation with human labels using correlation metrics. However, metrics like Krippendorff's α and Randolph's κ were originally designed to measure the reliability of human labeling, thus make assumptions about typical human labeling behavior, and these assumptions may not be applicable to machine generated labels. In this paper, we show how *relying on a single aggregate correlation score* can obscure fundamental differences between human labels and those from automatic evaluation, including LLM-as-a-Judge. Specifically, we demonstrate that when the proportion of samples with variation or uncertainty in human assigned labels is relatively high, machine labels (generated by automatic evaluation methods) may superficially appear to have similar or better correlation with the human majority label compared to the human-to-human (HH) correlation. This can create the illusion that labels from automatic evaluation approximates the human majority label. However, as the proportion of samples with consistent human labels increases, the correlation between machine and human labels fall well below HH correlation. Based on these findings, we first propose *stratifying data by human label uncertainty* to provide a more robust analysis of automatic evaluation performance. Second, recognizing that uncertainty and variation are inherent in perception-based human evaluations, such as those involving attitudes or preferences, we introduce a new metric -*binned Jensen-Shannon Divergence for perception* for such scenarios to better measure the effectiveness of automatic evaluations. Third, we present visualization techniques -- *perception charts*, to contextualize correlation measures appropriately and to show the strengths and limitations of automatic evaluation. We have open-sourced our analysis and visualization tools at https://github.com/amazon-science/BeyondCorrelation.

1207. RESuM: A Rare Event Surrogate Model for Physics Detector Design

链接：https://iclr.cc/virtual/2025/poster/28496 abstract： The experimental discovery of neutrinoless double-beta decay (NLDBD) would answer one of the most important questions in physics: Why is there more matter than antimatter in our universe? To maximize the chances of discovery, NLDBD experiments must optimize their detector designs to minimize the probability of background events contaminating the detector. Given that this probability is inherently low, design optimization either requires extremely costly simulations to generate sufficient background counts or contending with significant variance. In this work, we formalize this dilemma as a Rare Event Design (RED) problem: identifying optimal design parameters when the design metric to be minimized is inherently small. We then designed the Rare Event Surrogate Model (RESuM) for physics detector design optimization under RED conditions. RESuM uses a pre-trained Conditional Neural Process (CNP) model to incorporate additional prior knowledge into a Multi-Fidelity Gaussian Process model. We applied RESuM to optimize neutron shielding designs for the LEGEND NLDBD experiment, identifying an optimal design that reduces the neutron background by (66.5 ± 3.5)% while using only 3.3% of the computational resources compared to traditional methods. Given the prevalence of RED problems in other fields of physical sciences, especially in rare-event searches, the RESuM algorithm has broad potential for accelerating simulation-intensive applications.

1208. Reinforcement Learning for Control of Non-Markovian Cellular Population Dynamics

链接：https://iclr.cc/virtual/2025/poster/28960 abstract： Many organisms and cell types, from bacteria to cancer cells, exhibit a remarkable ability to adapt to fluctuating environments. Additionally, cells can leverage memory of past environments to better survive previously-encountered stressors. From a control perspective, this adaptability poses significant challenges in driving cell populations toward extinction, and is thus an open question with great clinical significance. In this work, we focus on drug dosing in cell populations exhibiting phenotypic plasticity. For specific dynamical models switching between resistant and susceptible states, exact solutions are known. However, when the underlying system parameters are unknown, and for complex memory-based systems, obtaining the optimal solution is currently intractable. To address this challenge, we apply reinforcement learning (RL) to identify informed dosing strategies to control cell populations evolving under novel non-Markovian dynamics. We find that model-free deep RL is able to recover exact solutions and control cell populations even in the presence of long-range temporal dynamics. To further test our approach in more realistic settings, we demonstrate performant RL-based control strategies in environments with dynamic memory strength.

1209. Policy Decorator: Model-Agnostic Online Refinement for Large Policy Model

链接：https://iclr.cc/virtual/2025/poster/28950 abstract： Recent advancements in robot learning have used imitation learning with large models and extensive demonstrations to develop effective policies. However, these models are often limited by the quantity quality, and diversity of demonstrations. This paper explores improving offline-trained imitation learning models through online interactions with the environment. We introduce Policy Decorator, which uses a model-agnostic residual policy to refine large imitation learning models during online interactions. By implementing controlled exploration strategies, Policy Decorator enables stable, sample-efficient online learning. Our evaluation spans eight tasks across two benchmarks—ManiSkill and Adroit—and involves two state-of-the-art imitation learning models (Behavior Transformer and Diffusion Policy). The results show Policy Decorator effectively improves the offline-trained policies and preserves the smooth motion of imitation learning models, avoiding the erratic behaviors of pure RL policies. See our project page for videos.

1210. Composable Interventions for Language Models

链接：https://iclr.cc/virtual/2025/poster/28014 abstract： Test-time interventions for language models can enhance factual accuracy, mitigate harmful outputs, and improve model efficiency without costly retraining. But despite a flood of new methods, different types of interventions are largely developing independently.In practice, multiple interventions must be applied sequentially to the same model, yet we lack standardized ways to study how interventions interact. We fill this gap by introducing composable interventions, a framework to study the effects of using multiple interventions on the same language models, featuring new metrics and a unified codebase. Using our framework, we conduct extensive experiments and compose popular methods from three emerging intervention categories---knowledge editing, model compression, and machine unlearning. Our results over 417 different compositions uncover meaningful interactions: compression hinders editing and unlearning, composing interventions hinges on their order of application, and popular general-purpose metrics are inadequate for assessing composability. Taken together, our findings showcase clear gaps in composability, suggesting a need for new multi-objective interventions.

1211. PRISM: Privacy-Preserving Improved Stochastic Masking for Federated Generative Models

链接：https://iclr.cc/virtual/2025/poster/30590 abstract： Despite recent advancements in federated learning (FL), the integration of generative models into FL has been limited due to challenges such as high communication costs and unstable training in heterogeneous data environments. To address these issues, we propose PRISM, a FL framework tailored for generative models that ensures (i) stable performance in heterogeneous data distributions and (ii) resource efficiency in terms of communication cost and final model size. The key of our method is to search for an optimal stochastic binary mask for a random network rather than updating the model weights, identifying a sparse subnetwork with high generative performance; i.e., a ``strong lottery ticket''. By communicating binary masks in a stochastic manner, PRISM minimizes communication overhead. This approach, combined with the utilization of maximum mean discrepancy (MMD) loss and a mask-aware dynamic moving average aggregation method (MADA) on the server side, facilitates stable and strong generative capabilities by mitigating local divergence in FL scenarios. Moreover, thanks to its sparsifying characteristic, PRISM yields a lightweight model without extra pruning or quantization, making it ideal for environments such as edge devices. Experiments on MNIST, FMNIST, CelebA, and CIFAR10 demonstrate that PRISM outperforms existing methods, while maintaining privacy with minimal communication costs. PRISM is the first to successfully generate images under challenging non-IID and privacy-preserving FL environments on complex datasets, where previous methods have struggled.

1212. Graph-based Document Structure Analysis

链接：https://iclr.cc/virtual/2025/poster/30318 abstract： When reading a document, glancing at the spatial layout of a document is an initial step to understand it roughly. Traditional document layout analysis (DLA) methods, however, offer only a superficial parsing of documents, focusing on basic instance detection and often failing to capture the nuanced spatial and logical relationships between instances. These limitations hinder DLA-based models from achieving a gradually deeper comprehension akin to human reading. In this work, we propose a novel graph-based Document Structure Analysis (gDSA) task. This task requires that model not only detects document elements but also generates spatial and logical relations in form of a graph structure, allowing to understand documents in a holistic and intuitive manner. For this new task, we construct a relation graph-based document structure analysis dataset(GraphDoc) with 80K document images and 4.13M relation annotations, enabling training models to complete multiple tasks like reading order, hierarchical structures analysis, and complex inter-element relationship inference. Furthermore, a document relation graph generator (DRGG) is proposed to address the gDSA task, which achieves performance with 57.6% at mAP_g@0.5 for a strong benchmark baseline on this novel task and dataset. We hope this graphical representation of document structure can mark an innovative advancement in document structure analysis and understanding. The new dataset and code will be made publicly available.

1213. SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding

链接：https://iclr.cc/virtual/2025/poster/32102 abstract： Multimodal large language models (MLLMs) have recently shown great progress in text-rich image understanding, yet they still struggle with complex, multi-page visually-rich documents. Traditional methods using document parsers for retrieval-augmented generation suffer from performance and efficiency limitations, while directly presenting all pages to MLLMs leads to inefficiencies, especially with lengthy ones. In this work, we present a novel framework named Self-Visual Retrieval-Augmented Generation (SV-RAG), which can broaden horizons of any MLLM to support long-document understanding. We demonstrate that MLLMs themselves can be an effective multimodal retriever to fetch relevant pages and then answer user questions based on these pages. SV-RAG is implemented with two specific MLLM adapters, one for evidence page retrieval and the other for question answering. Empirical results show state-of-the-art performance on public benchmarks, demonstrating the effectiveness of SV-RAG.

1214. LICO: Large Language Models for In-Context Molecular Optimization

链接：https://iclr.cc/virtual/2025/poster/27698 abstract： Optimizing black-box functions is a fundamental problem in science and engineering. To solve this problem, many approaches learn a surrogate function that estimates the underlying objective from limited historical evaluations. Large Language Models (LLMs), with their strong pattern-matching capabilities via pretraining on vast amounts of data, stand out as a potential candidate for surrogate modeling. However, directly prompting a pretrained language model to produce predictions is not feasible in many scientific domains due to the scarcity of domain-specific data in the pretraining corpora and the challenges of articulating complex problems in natural language. In this work, we introduce LICO, a general-purpose model that extends arbitrary base LLMs for black-box optimization, with a particular application to the molecular domain. To achieve this, we equip the language model with a separate embedding layer and prediction layer, and train the model to perform in-context predictions on a diverse set of functions defined over the domain. Once trained, LICO can generalize to unseen molecule properties simply via in-context prompting. LICO performs competitively on PMO, a challenging molecular optimization benchmark comprising 23 objective functions, and achieves state-of-the-art performance on its low-budget version PMO-1K.

1215. Bisimulation Metric for Model Predictive Control

链接：https://iclr.cc/virtual/2025/poster/30370 abstract： Model-based reinforcement learning (MBRL) has shown promise for improving sample efficiency and decision-making in complex environments. However, existing methods face challenges in training stability, robustness to noise, and computational efficiency. In this paper, we propose Bisimulation Metric for Model Predictive Control (BS-MPC), a novel approach that incorporates bisimulation metric loss in its objective function to directly optimize the encoder. This optimization enables the learned encoder to extract intrinsic information from the original state space while discarding irrelevant details. BS-MPC improves training stability, robustness against input noise, and computational efficiency by reducing training time. We evaluate BS-MPC on both continuous control and image-based tasks from the DeepMind Control Suite, demonstrating superior performance and robustness compared to state-of-the-art baseline methods.

1216. Causal Identification for Complex Functional Longitudinal Studies

链接：https://iclr.cc/virtual/2025/poster/30722 abstract： Real-time monitoring in modern medical research introduces functional longitudinal data, characterized by continuous-time measurements of outcomes, treatments, and confounders. This complexity leads to uncountably infinite treatment-confounder feedbacks, which traditional causal inference methodologies cannot handle. Inspired by the coarsened data framework, we adopt stochastic process theory, measure theory, and net convergence to propose a nonparametric causal identification framework. This framework generalizes classical g-computation, inverse probability weighting, and doubly robust formulas, accommodating time-varying outcomes subject to mortality and censoring for functional longitudinal data. We examine our framework through Monte Carlo simulations. Our approach addresses significant gaps in current methodologies, providing a solution for functional longitudinal data and paving the way for future estimation work in this domain.

1217. Geometry of Lightning Self-Attention: Identifiability and Dimension

链接：https://iclr.cc/virtual/2025/poster/29285 abstract： We consider function spaces defined by self-attention networks without normalization, and theoretically analyze their geometry. Since these networks are polynomial, we rely on tools from algebraic geometry. In particular, we study the identifiability of deep attention by providing a description of the generic fibers of the parametrization for an arbitrary number of layers and, as a consequence, compute the dimension of the function space. Additionally, for a single-layer model, we characterize the singular and boundary points. Finally, we formulate a conjectural extension of our results to normalized self-attention networks, prove it for a single layer, and numerically verify it in the deep case.

1218. Taming Overconfidence in LLMs: Reward Calibration in RLHF

链接：https://iclr.cc/virtual/2025/poster/28542 abstract： Language model calibration refers to the alignment between the confidence of the model and the actual performance of its responses.While previous studies point out the overconfidence phenomenon in Large Language Models (LLMs) and show that LLMs trained with Reinforcement Learning from Human Feedback (RLHF) are overconfident with a more sharpened output probability, in this study, we reveal that RLHF tends to lead models to express verbalized overconfidence in their own responses. We investigate the underlying cause of this overconfidence and demonstrate that reward models used for Proximal Policy Optimization (PPO) exhibit inherent biases towards high-confidence scores regardless of the actual quality of responses. Building upon this insight, we propose two PPO variants: PPO-M: $\underline{PPO}$ with Calibrated Reward $\underline{M}$odeling and PPO-C: $\underline{PPO}$ with Calibrated Reward $\underline{C}$alculation. PPO-M integrates explicit confidence scores in reward model training, which calibrates reward modelsto better capture the alignment between response quality and verbalized confidence. PPO-C adjusts the reward score during PPO based on the difference between the current reward and the moving average of past rewards. Both PPO-M and PPO-C can be seamlessly integrated into the current PPO pipeline and do not require additional golden labels. We evaluate our methods on both $\texttt{Llama3-8B}$ and $\texttt{Mistral-7B}$ across six diverse datasets including multiple-choice and open-ended generation. Experiment results demonstrate that both of our methods can reduce calibration error and maintain performance comparable to standard PPO. We further show that they do not compromise model capabilities in open-ended conversation settings.

1219. Fourier Head: Helping Large Language Models Learn Complex Probability Distributions

链接：https://iclr.cc/virtual/2025/poster/30999 abstract： As the quality of large language models has improved, there has been increased interest in using them to model non-linguistic tokens. For example, the Decision Transformer recasts agentic decision making as a sequence modeling problem, using a decoder-only LLM to model the distribution over the discrete action space for an Atari agent. However, when adapting LLMs to non-linguistic domains, it remains unclear if softmax over discrete bins captures the continuous structure of the tokens and the potentially complex distributions needed for high quality token generation. We introduce a neural network layer, constructed using Fourier series, which we can easily substitute for any linear layer if we want the outputs to have a more continuous structure. We perform extensive analysis on synthetic datasets, as well as on large-scale decision making and time series forecasting tasks. We also provide theoretical evidence that this layer can better learn signal from data while ignoring high-frequency noise. All of our results support the effectiveness of our proposed Fourier head in scenarios where the underlying data distribution has a natural continuous structure. For example, the Fourier head improves a Decision Transformer agent's returns across four benchmark Atari games by as much as 377\%, and increases a state-of-the-art times series foundation model's forecasting performance by 3.5\% across 20 benchmarks unseen during training.We release our implementation at https://nategillman.com/fourier-head

1220. Mutual Effort for Efficiency: A Similarity-based Token Pruning for Vision Transformers in Self-Supervised Learning

链接：https://iclr.cc/virtual/2025/poster/30276 abstract： Self-supervised learning (SSL) offers a compelling solution to the challenge of extensive labeled data requirements in traditional supervised learning.With the proven success of Vision Transformers (ViTs) in supervised tasks, there is increasing interest in adapting them for SSL frameworks. However, the high computational demands of SSL pose substantial challenges, particularly on resource-limited platforms like edge devices, despite its ability to achieve high accuracy without labeled data.Recent studies in supervised learning have shown that token pruning can reduce training costs by removing less informative tokens without compromising accuracy. However, SSL’s dual-branch encoders make traditional single-branch pruning strategies less effective, as they fail to account for the critical cross-branch similarity information, leading to reduced accuracy in SSL.To this end, we introduce SimPrune, a novel token pruning strategy designed for ViTs in SSL. SimPrune leverages cross-branch similarity information to efficiently prune tokens, retaining essential semantic information across dual branches. Additionally, we incorporate a difficulty-aware pruning strategy to further enhance SimPrune's effectiveness.Experimental results show that our proposed approach effectively reduces training computation while maintaining accuracy. Specifically, our approach offers 24\% savings in training costs compared to SSL baseline, without sacrificing accuracy.

1221. DiffusionGuard: A Robust Defense Against Malicious Diffusion-based Image Editing

链接：https://iclr.cc/virtual/2025/poster/30694 abstract： Recent advances in diffusion models have introduced a new era of text-guided image manipulation, enabling users to create realistic edited images with simple textual prompts. However, there is significant concern about the potential misuse of these methods, especially in creating misleading or harmful content. Although recent defense strategies, which introduce imperceptible adversarial noise to induce model failure, have shown promise, they remain ineffective against more sophisticated manipulations, such as editing with a mask. In this work, we propose DiffusionGuard, a robust and effective defense method against unauthorized edits by diffusion-based image editing models, even in challenging setups. Through a detailed analysis of these models, we introduce a novel objective that generates adversarial noise targeting the early stage of the diffusion process. This approach significantly improves the efficiency and effectiveness of adversarial noises. We also introduce a mask-augmentation technique to enhance robustness against various masks during test time. Finally, we introduce a comprehensive benchmark designed to evaluate the effectiveness and robustness of methods in protecting against privacy threats in realistic scenarios. Through extensive experiments, we show that our method achieves stronger protection and improved mask robustness with lower computational costs compared to the strongest baseline. Additionally, our method exhibits superior transferability and better resilience to noise removal techniques compared to all baseline methods. Our source code is publicly available at https://choi403.github.io/diffusionguard.

1222. TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness for Generalist Robotic Policies

链接：https://iclr.cc/virtual/2025/poster/29130 abstract： Although large vision-language-action (VLA) models pretrained on extensive robot datasets offer promising generalist policies for robotic learning, they still struggle with spatial-temporal dynamics in interactive robotics, making them less effective in handling complex tasks, such as manipulation. In this work, we introduce visual trace prompting, a simple yet effective approach to facilitate VLA models' spatial-temporal awareness for action prediction by encoding state-action trajectories visually. We develop a new TraceVLA model by finetuningOpenVLA on our own collected dataset of 150K robot manipulation trajectories using visual trace prompting. Evaluations of TraceVLA across 137 configurations in SimplerEnv and 4 tasks on a physical WidowX robot demonstrate state-of-the-art performance, outperforming OpenVLA by 10% on SimplerEnv and 3.5x on real-robot tasks and exhibiting robust generalization across diverse embodiments and scenarios. To further validate the effectiveness and generality of our method, we present a compact VLA model based on 4B Phi-3-Vision, pretrained on the Open-X-Embodiment and finetuned on our dataset, rivals the 7B OpenVLA baseline while significantly improving inference efficiency.

1223. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think

链接：https://iclr.cc/virtual/2025/poster/30467 abstract： Recent studies have shown that the denoising process in (generative) diffusion models can induce meaningful (discriminative) representations inside the model, though the quality of these representations still lags behind those learned through recent self-supervised learning methods. We argue that one main bottleneck in training large-scale diffusion models for generation lies in effectively learning these representations. Moreover, training can be made easier by incorporating high-quality external visual representations, rather than relying solely on the diffusion models to learn them independently. We study this by introducing a straightforward regularization called REPresentation Alignment (REPA), which aligns the projections of noisy input hidden states in denoising networks with clean image representations obtained from external, pretrained visual encoders. The results are striking: our simple strategy yields significant improvements in both training efficiency and generation quality when applied to popular diffusion and flow-based transformers, such as DiTs and SiTs. For instance, our method can speed up SiT training by over 17.5$\times$, matching the performance (without classifier-free guidance) of a SiT-XL model trained for 7M steps in less than 400K steps. In terms of final generation quality, our approach achieves state-of-the-art results of FID=1.42 using classifier-free guidance with the guidance interval.

1224. RevisEval: Improving LLM-as-a-Judge via Response-Adapted References

链接：https://iclr.cc/virtual/2025/poster/31178 abstract： With significant efforts in recent studies, LLM-as-a-Judge has become a cost-effective alternative to human evaluation for assessing text generation quality in a wide range of tasks. However, there still remains a reliability gap between LLM-as-a-Judge and human evaluation. One important reason is the lack of guided oracles in the evaluation process. Motivated by the role of reference pervasively used in classic text evaluation, we introduce RevisEval, a novel text generation evaluation paradigm via the response-adapted references. RevisEval is driven by the key observation that an ideal reference should maintain the necessary relevance to the response to be evaluated. Specifically, RevisEval leverages the text revision capabilities of large language models (LLMs) to adaptively revise the response, then treat the revised text as the reference (response-adapted reference) for the subsequent evaluation. Extensive experiments demonstrate that RevisEval outperforms traditional reference-free and reference-based evaluation paradigms that use LLM-as-a-Judge across NLG tasks and open-ended instruction-following tasks. More importantly, our response-adapted references can further boost the classical text metrics, e.g., BLEU and BERTScore, compared to traditional references and even rival the LLM-as-a-Judge. A detailed analysis is also conducted to confirm RevisEval's effectiveness in bias reduction, the impact of inference cost, and reference relevance.

1225. Efficient Model Editing with Task-Localized Sparse Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/29556 abstract： Task arithmetic has emerged as a promising approach for editing models by representing task-specific knowledge as composable task vectors. However, existing methods rely on network linearization to derive task vectors, leading to computational bottlenecks during training and inference. Moreover, linearization alone does not ensure weight disentanglement, the key property that enables conflict-free composition of task vectors. To address this, we propose TaLoS which allows to build sparse task vectors with minimal interference without requiring explicit linearization and sharing information across tasks. We find that pre-trained models contain a subset of parameters with consistently low gradient sensitivity across tasks, and that sparsely updating only these parameters allows for promoting weight disentanglement during fine-tuning. Our experiments prove that TaLoS improves training and inference efficiency while outperforming current methods in task addition and negation. By enabling modular parameter editing, our approach fosters practical deployment of adaptable foundation models in real-world applications.

1226. Minimalistic Predictions for Online Class Constraint Scheduling

链接：https://iclr.cc/virtual/2025/poster/28651 abstract： We consider online scheduling with class constraints. That is, we are given m machines, each with k class slots. Upon receiving a job j with class c_j, an algorithm needs to allocate j on some machine i. The goal is to minimize the makespan while not assigning more than k different classes onto each machine.While the offline case is well understood and even (E)PTAS results are known [Jansen, Lassota, Maack SPAA'20, Chen Jansen Luo Zhang COCOA'16], the online case admits strong impossibility results in classical competitive analysis [Epstein, Lassota, Levin, Maack, Rohwedder STACS'22].We overcome these daunting results by investigating the problem in a learning-augmented setting where an algorithm can access possibly erroneous predictions. We present new algorithms with competitive ratios independent of m and tight lower bounds for several classical and problem-specific prediction models. We thereby give a structured overview of what additional information helps in the design of better scheduling algorithms.

1227. Linear combinations of latents in generative models: subspaces and beyond

链接：https://iclr.cc/virtual/2025/poster/28429 abstract： Sampling from generative models has become a crucial tool for applications like data synthesis and augmentation. Diffusion, Flow Matching and Continuous Normalizing Flows have shown effectiveness across various modalities, and rely on latent variables for generation. For experimental design or creative applications that require more control over the generation process, it has become common to manipulate the latent variable directly. However, existing approaches for performing such manipulations (e.g. interpolation or forming low-dimensional representations) only work well in special cases or are network or data-modality specific. We propose Linear combinations of Latent variables (LOL) as a general-purpose method to form linear combinations of latent variables that adhere to the assumptions of the generative model. As LOL is easy to implement and naturally addresses the broader task of forming any linear combinations, e.g. the construction of subspaces of the latent space, LOL dramatically simplifies the creation of expressive low-dimensional representations of high-dimensional objects.

1228. XAIguiFormer: explainable artificial intelligence guided transformer for brain disorder identification

链接：https://iclr.cc/virtual/2025/poster/30644 abstract： EEG-based connectomes offer a low-cost and portable method to identify brain disorders using deep learning. With the growing interest in model interpretability and transparency, explainable artificial intelligence (XAI) is widely applied to understand the decision of deep learning models. However, most research focuses solely on interpretability analysis based on the insights from XAI, overlooking XAI's potential to improve model performance. To bridge this gap, we propose a dynamical-system-inspired architecture, XAI guided transformer (XAIguiFormer), where XAI not only provides explanations but also contributes to enhancing the transformer by refining the originally coarse information in self-attention mechanism to capture more relevant dependency relationships. In order not to damage the connectome's topological structure, the connectome tokenizer treats the single-band graphs as atomic tokens to generate a sequence in the frequency domain. To address the limitations of conventional positional encoding in understanding the frequency and mitigating the individual differences, we integrate frequency and demographic information into tokens via a rotation matrix, resulting in a richly informative representation. Our experiment demonstrates that XAIguiFormer achieves superior performance over all baseline models. In addition, XAIguiFormer provides valuable interpretability through visualization of the frequency band importance. Our code is available at https://github.com/HanningGuo/XAIguiFormer.

1229. Benchmarking Predictive Coding Networks -- Made Simple

链接：https://iclr.cc/virtual/2025/poster/28118 abstract： In this work, we tackle the problems of efficiency and scalability for predictive coding networks (PCNs) in machine learning. To do so, we propose a library that focuses on performance and simplicity, and use it to implement a large set of standard benchmarks for the community to use for their experiments. As most works in the field propose their own tasks and architectures, do not compare one against each other, and focus on small-scale tasks, a simple and fast open-source library, and a comprehensive set of benchmarks, would address all of these concerns. Then, we perform extensive tests on such benchmarks using both existing algorithms for PCNs, as well as adaptations of other methods popular in the bio-plausible deep learning community. All of this has allowed us to (i) test architectures much larger than commonly used in the literature, on more complex datasets; (ii) reach new state-of-the-art results in all of the tasks and dataset provided; (iii) clearly highlight what the current limitations of PCNs are, allowing us to state important future research directions. With the hope of galvanizing community efforts towards one of the main open problems in the field, scalability, we will release the code, tests, and benchmarks.

1230. Shifting the Paradigm: A Diffeomorphism Between Time Series Data Manifolds for Achieving Shift-Invariancy in Deep Learning

链接：https://iclr.cc/virtual/2025/poster/28398 abstract： Deep learning models lack shift invariance, making them sensitive to input shifts that cause changes in output. While recent techniques seek to address this for images, our findings show that these approaches fail to provide shift-invariance in time series, where the data generation mechanism is more challenging due to the interaction of low and high frequencies. Worse, they also decrease performance across several tasks. In this paper, we propose a novel differentiable bijective function that maps samples from their high-dimensional data manifold to another manifold of the same dimension, without any dimensional reduction. Our approach guarantees that samples---when subjected to random shifts---are mapped to a unique point in the manifold while preserving all task-relevant information without loss. We theoretically and empirically demonstrate that the proposed transformation guarantees shift-invariance in deep learning models without imposing any limits to the shift. Our experiments on six time series tasks with state-of-the-art methods show that our approach consistently improves the performance while enabling models to achieve complete shift-invariance without modifying or imposing restrictions on the model's topology. The source code is available on GitHub.

1231. Data Distillation for extrapolative protein design through exact preference optimization

链接：https://iclr.cc/virtual/2025/poster/27969 abstract： The goal of protein design typically involves increasing fitness (extrapolating) beyond what is seen during training (e.g., towards higher stability, stronger binding affinity, etc.). State-of-the-art methods assume that one can safely steer proteins towards such extrapolated regions by learning from pairs alone. We hypothesize that noisy training pairs are not sufficiently informative to capture the fitness gradient and that models learned from pairs specifically may fail to capture three-way relations important for search, e.g., how two alternatives fair relative to a seed. Building on the success of preference alignment models in large language models, we introduce a progressive search method for extrapolative protein design by directly distilling into the model relevant triplet relations. We evaluated our model's performance in designing AAV and GFP proteins and demonstrated that the proposed framework significantly improves effectiveness in extrapolation tasks.

1232. Risk-Controlling Model Selection via Guided Bayesian Optimization

链接：https://iclr.cc/virtual/2025/poster/31460 abstract： Adjustable hyperparameters of machine learning models typically impact various key trade-offs such as accuracy, fairness, robustness, or inference cost. Our goal in this paper is to find a configuration that adheres to user-specified limits on certain risks while being useful with respect to other conflicting metrics. We solve this by combining Bayesian Optimization (BO) with rigorous risk-controlling procedures, where our core idea is to steer BO towards an efficient testing strategy. Our BO method identifies a set of Pareto optimal configurations residing in a designated region of interest. The resulting candidates are statistically verified, and the best-performing configuration is selected with guaranteed risk levels. We demonstrate the effectiveness of our approach on a range of tasks with multiple desiderata, including low error rates, equitable predictions, handling spurious correlations, managing rate and distortion in generative models, and reducing computational costs.

1233. On Scaling Up 3D Gaussian Splatting Training

链接：https://iclr.cc/virtual/2025/poster/28300 abstract： 3D Gaussian Splatting (3DGS) is increasingly popular for 3D reconstruction due to its superior visual quality and rendering speed. However, 3DGS training currently occurs on a single GPU, limiting its ability to handle high-resolution and large-scale 3D reconstruction tasks due to memory constraints. We introduce Grendel, a distributed system designed to partition 3DGS parameters and parallelize computation across multiple GPUs. As each Gaussian affects a small, dynamic subset of rendered pixels, Grendel employs sparse all-to-all communication to transfer the necessary Gaussians to pixel partitions and performs dynamic load balancing. Unlike existing 3DGS systems that train using one camera view image at a time, Grendel supports batched training with multiple views. We explore various optimization hyperparameter scaling strategies and find that a simple sqrt(batch-size) scaling rule is highly effective. Evaluations using large-scale, high-resolution scenes show that Grendel enhances rendering quality by scaling up 3DGS parameters across multiple GPUs. On the 4K ``Rubble'' dataset, we achieve a test PSNR of 27.28 by distributing 40.4 million Gaussians across 16 GPU, compared to a PSNR of 26.28 using 11.2 million Gaussians on a single GPU. Grendel is an open-source project available at: https://github.com/nyu-systems/Grendel-GS

1234. Coreset Selection via Reducible Loss in Continual Learning

链接：https://iclr.cc/virtual/2025/poster/28476 abstract： Rehearsal-based continual learning (CL) aims to mitigate catastrophic forgetting by maintaining a subset of samples from previous tasks and replaying them. The rehearsal memory can be naturally constructed as a coreset, designed to form a compact subset that enables training with performance comparable to using the full dataset. The coreset selection task can be formulated as bilevel optimization that solves for the subset to minimize the outer objective of the learning task. Existing methods primarily rely on inefficient probabilistic sampling or local gradient-based scoring to approximate sample importance through an iterative process that can be susceptible to ambiguity or noise. Specifically, non-representative samples like ambiguous or noisy samples are difficult to learn and incur high loss values even when training on the full dataset. However, existing methods relying on local gradient tend to highlight these samples in an attempt to minimize the outer loss, leading to a suboptimal coreset. To enhance coreset selection, especially in CL where high-quality samples are essential, we propose a coreset selection method that measures sample importance using reducible loss (ReL) that quantifies the impact of adding a sample to model performance. By leveraging ReL and a process derived from bilevel optimization, we identify and retain samples that yield the highest performance gain. They are shown to be informative and representative. Furthermore, ReL requires only forward computation, making it significantly more efficient than previous methods. To better apply coreset selection in CL, we extend our method to address key challenges such as task interference, streaming data, and knowledge distillation. Experiments on data summarization and continual learning demonstrate the effectiveness and efficiency of our approach.

1235. Generator Matching: Generative modeling with arbitrary Markov processes

链接：https://iclr.cc/virtual/2025/poster/29636 abstract： We introduce Generator Matching, a modality-agnostic framework for generative modeling using arbitrary Markov processes. Generators characterize the infinitesimal evolution of a Markov process, which we leverage for generative modeling in a similar vein to flow matching: we construct conditional generators which generate single data points, then learn to approximate the marginal generator which generates the full data distribution. We show that Generator Matching unifies various generative modeling methods, including diffusion models, flow matching and discrete diffusion models. Furthermore, it expands the design space to new and unexplored Markov processes such as jump processes. Finally, Generator Matching enables the construction of superpositions of Markov generative models and enables the construction of multimodal models in a rigorous manner. We empirically validate our method on image and multimodal generation, e.g. showing that superposition with a jump process improves performance.

1236. A Percolation Model of Emergence: Analyzing Transformers Trained on a Formal Language

链接：https://iclr.cc/virtual/2025/poster/31234 abstract： Increase in data, size, or compute can lead to sudden learning of specific capabilities by a neural network---a phenomenon often called "emergence". Beyond scientific understanding, establishing the causal factors underlying such emergent capabilities is crucial to enable risk regulation frameworks for AI. In this work, we seek inspiration from study of emergent properties in other fields and propose a phenomenological definition for the concept in the context of neural networks. Our definition implicates the acquisition of general regularities underlying the data-generating process as a cause of sudden performance growth for specific, narrower tasks. We empirically investigate this definition by proposing an experimental system grounded in a context-sensitive formal language, and find that Transformers trained to perform tasks on top of strings from this language indeed exhibit emergent capabilities. Specifically, we show that once the language's underlying grammar and context-sensitivity inducing regularities are learned by the model, performance on narrower tasks suddenly begins to improve. We then analogize our network's learning dynamics with the process of percolation on a bipartite graph, establishing a formal phase transition model that predicts the shift in the point of emergence observed in our experiments when intervening on the data regularities. Overall, our experimental and theoretical frameworks yield a step towards better defining, characterizing, and predicting emergence in neural networks.

1237. PhyloVAE: Unsupervised Learning of Phylogenetic Trees via Variational Autoencoders

链接：https://iclr.cc/virtual/2025/poster/29234 abstract： Learning informative representations of phylogenetic tree structures is essential for analyzing evolutionary relationships. Classical distance-based methods have been widely used to project phylogenetic trees into Euclidean space, but they are often sensitive to the choice of distance metric and may lack sufficient resolution. In this paper, we introduce phylogenetic variational autoencoders (PhyloVAEs), an unsupervised learning framework designed for representation learning and generative modeling of tree topologies. Leveraging an efficient encoding mechanism inspired by autoregressive tree topology generation, we develop a deep latent-variable generative model that facilitates fast, parallelized topology generation. PhyloVAE combines this generative model with a collaborative inference model based on learnable topological features, allowing for high-resolution representations of phylogenetic tree samples. Extensive experiments demonstrate PhyloVAE's robust representation learning capabilities and fast generation of phylogenetic tree topologies.

1238. Underdamped Diffusion Bridges with Applications to Sampling

链接：https://iclr.cc/virtual/2025/poster/29713 abstract： We provide a general framework for learning diffusion bridges that transport prior to target distributions. It includes existing diffusion models for generative modeling, but also underdamped versions with degenerate diffusion matrices, where the noise only acts in certain dimensions. Extending previous findings, our framework allows to rigorously show that score-matching in the underdamped case is indeed equivalent to maximizing a lower bound on the likelihood. Motivated by superior convergence properties and compatibility with sophisticated numerical integration schemes of underdamped stochastic processes, we propose underdamped diffusion bridges, where a general density evolution is learned rather than prescribed by a fixed noising process. We apply our method to the challenging task of sampling from unnormalized densities without access to samples from the target distribution. Across a diverse range of sampling problems, our approach demonstrates state-of-the-art performance, notably outperforming alternative methods, while requiring significantly fewer discretization steps and almost no hyperparameter tuning.

1239. Simulating Training Dynamics to Reconstruct Training Data from Deep Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29225 abstract： Whether deep neural networks (DNNs) memorize the training data is a fundamental open question in understanding deep learning. A direct way to verify the memorization of DNNs is to reconstruct training data from DNNs’ parameters. Since parameters are gradually determined by data throughout training, characterizing training dynamics is important for reconstruction. Pioneering works rely on the linear training dynamics of shallow NNs with large widths, but cannot be extended to more practical DNNs which have non-linear dynamics. We propose Simulation of training Dynamics (SimuDy) to reconstruct training data from DNNs. Specifically, we simulate the training dynamics by training the model from the initial parameters with a dummy dataset, then optimize this dummy dataset so that the simulated dynamics reach the same final parameters as the true dynamics. By incorporating dummy parameters in the simulated dynamics, SimuDy effectively describes non-linear training dynamics. Experiments demonstrate that SimuDy significantly outperforms previous approaches when handling non-linear training dynamics, and for the first time, most training samples can be reconstructed from a trained ResNet’s parameters.

1240. Sequential Controlled Langevin Diffusions

链接：https://iclr.cc/virtual/2025/poster/28991 abstract： An effective approach for sampling from unnormalized densities is based on the idea of gradually transporting samples from an easy prior to the complicated target distribution. Two popular methods are (1) Sequential Monte Carlo (SMC), where the transport is performed through successive annealed densities via prescribed Markov chains and resampling steps, and (2) recently developed diffusion-basedsampling methods, where a learned dynamical transport is used. Despite the common goal, both approaches have different, often complementary, advantages and drawbacks. The resampling steps in SMC allow focusing on promising regions of the space, often leading to robust performance. While the algorithm enjoys asymptotic guarantees, the lack of flexible, learnable transitions can lead to slow convergence. On the other hand, diffusion-based samplers are learned and can potentially better adapt themselves to the target at hand, yet often suffer from training instabilities. In this work, we present a principled framework for combining SMC with diffusion-based samplers by viewing both methods in continuous time and considering measures on path space. This culminates in the new Sequential Controlled Langevin Diffusion (SCLD) sampling method, which is able to utilize the benefits of both methods and reaches improved performance on multiple benchmark problems, in many cases using only 10% of the training budget of previous diffusion-based samplers.

1241. Interpreting Global Perturbation Robustness of Image Models using Axiomatic Spectral Importance Decomposition

链接：https://iclr.cc/virtual/2025/poster/31471 abstract： Perturbation robustness evaluates the vulnerabilities of models, arising from a variety of perturbations, such as data corruptions and adversarial attacks. Understanding the mechanisms of perturbation robustness is critical for global interpretability. We present a model-agnostic, global mechanistic interpretability method to interpret the perturbation robustness of image models. This research is motivated by two key aspects. First, previous global interpretability works, in tandem with robustness benchmarks, eg. mean corruption error (mCE), are not designed to directly interpret the mechanisms of perturbation robustness within image models. Second, we notice that the spectral signal-to-noise ratios (SNR) of perturbed natural images exponentially decay over the frequency. This power-law-like decay implies that: Low-frequency signals are generally more robust than high-frequency signals -- yet high classification accuracy can not be achieved by low-frequency signals alone. By applying Shapley value theory, our method axiomatically quantifies the predictive powers of robust features and non-robust features within an information theory framework. Our method, dubbed as I-ASIDE (Image Axiomatic Spectral Importance Decomposition Explanation), provides a unique insight into model robustness mechanisms. We conduct extensive experiments over a variety of vision models pre-trained on ImageNet, including both convolutional neural networks (eg. AlexNet, VGG, GoogLeNet/Inception-v1, Inception-v3, ResNet, SqueezeNet, RegNet, MnasNet, MobileNet, EfficientNet, etc.) and vision transformers (eg. ViT, Swin Transformer, and MaxViT), to show that I-ASIDE can not only measure the perturbation robustness but also provide interpretations of its mechanisms.

1242. Solving Inverse Problems with Model Mismatch using Untrained Neural Networks within Model-based Architectures

链接：https://iclr.cc/virtual/2025/poster/31484 abstract： Model-based deep learning methods such as loop unrolling (LU) and deep equilibrium model (DEQ) extensions offer outstanding performance in solving inverse problems (IP). These methods unroll the optimization iterations into a sequence of neural networks that in effect learn a regularization function from data. While these architectures are currently state-of-the-art in numerous applications, their success heavily relies on the accuracy of the forward model. This assumption can be limiting in many physical applications due to model simplifications or uncertainties in the apparatus. To address forward model mismatch, we introduce an untrained forward model residual block within the model-based architecture to match the data consistency in the measurement domain for each instance. We propose two variants in well-known model-based architectures (LU and DEQ) and prove convergence under mild conditions. Our approach offers a unified solution that is less parameter-sensitive, requires no additional data, and enables simultaneous fitting of the forward model and reconstruction in a single pass, benefiting both linear and nonlinear inverse problems. The experiments show significant quality improvement in removing artifacts and preserving details across three distinct applications, encompassing both linear and nonlinear inverse problems. Moreover, we highlight reconstruction effectiveness in intermediate steps and showcase robustness to random initialization of the residual block and a higher number of iterations during evaluation.

1243. GraphArena: Evaluating and Exploring Large Language Models on Graph Computation

链接：https://iclr.cc/virtual/2025/poster/29278 abstract： The ``arms race'' of Large Language Models (LLMs) demands new benchmarks to examine their progresses. In this paper, we introduce GraphArena, a benchmarking tool designed to evaluate LLMs on real-world graph computational problems. It offers a suite of four polynomial-time tasks (e.g., Shortest Distance) and six NP-complete challenges (e.g., Traveling Salesman Problem). GraphArena features a rigorous evaluation framework that classifies LLM outputs as correct, suboptimal (feasible but not optimal), hallucinatory (properly formatted but infeasible), or missing. Evaluation of over 10 LLMs reveals that even top-performing LLMs struggle with larger, more complex graph problems and exhibit hallucination issues. We further explore four potential solutions to address this issue and improve LLMs on graph computation, including chain-of-thought prompting, instruction tuning, code writing, and scaling test-time compute, each demonstrating unique strengths and limitations. GraphArena complements the existing LLM benchmarks and is open-sourced at https://github.com/squareRoot3/GraphArena.

1244. GDrag:Towards General-Purpose Interactive Editing with Anti-ambiguity Point Diffusion

链接：https://iclr.cc/virtual/2025/poster/30776 abstract： Recent interactive point-based image manipulation methods have gained considerable attention for being user-friendly. However, these methods still face two types of ambiguity issues that can lead to unsatisfactory outcomes, namely, intention ambiguity which misinterprets the purposes of users, and content ambiguity where target image areas are distorted by distracting elements. To address these issues and achieve general-purpose manipulations, we propose a novel task-aware, training-free framework called GDrag. Specifically, GDrag defines a taxonomy of atomic manipulations, which can be parameterized and combined unitedly to represent complex manipulations, thereby reducing intention ambiguity. Furthermore, GDrag introduces two strategies to mitigate content ambiguity, including an anti-ambiguity dense trajectory calculation method (ADT) and a self-adaptive motion supervision method (SMS). Given an atomic manipulation, ADT converts the sparse user-defined handle points into a dense point set by selecting their semantic and geometric neighbors, and calculates the trajectory of the point set. Unlike previous motion supervision methods relying on a single global scale for low-rank adaption, SMS jointly optimizes point-wise adaption scales and latent feature biases. These two methods allow us to model fine-grained target contexts and generate precise trajectories. As a result, GDrag consistently produces precise and appealing results in different editing tasks. Extensive experiments on the challenging DragBench dataset demonstrate that GDrag outperforms state-of-the-art methods significantly. The code of GDrag will be released upon acceptance.

1245. Find A Winning Sign: Sign Is All We Need to Win the Lottery

链接：https://iclr.cc/virtual/2025/poster/29056 abstract： The Lottery Ticket Hypothesis (LTH) posits the existence of a sparse subnetwork (a.k.a. winning ticket) that can generalize comparably to its over-parameterized counterpart when trained from scratch.The common approach to finding a winning ticket is to preserve the original strong generalization through Iterative Pruning (IP) and transfer information useful for achieving the learned generalization by applying the resulting sparse mask to an untrained network.However, existing IP methods still struggle to generalize their observations beyond ad-hoc initialization and small-scale architectures or datasets, or they bypass these challenges by applying their mask to trained weights instead of initialized ones.In this paper, we demonstrate that the parameter sign configuration plays a crucial role in conveying useful information for generalization to any randomly initialized network.Through linear mode connectivity analysis, we observe that a sparse network trained by an existing IP method can retain its basin of attraction if its parameter signs and normalization layer parameters are preserved.To take a step closer to finding a winning ticket, we alleviate the reliance on normalization layer parameters by preventing high error barriers along the linear path between the sparse network trained by our method and its counterpart with initialized normalization layer parameters.Interestingly, across various architectures and datasets, we observe that any randomly initialized network can be optimized to exhibit low error barriers along the linear path to the sparse network trained by our method by inheriting its sparsity and parameter sign information, potentially achieving performance comparable to the original.The code is available at https://github.com/JungHunOh/AWS_ICLR2025.git.

1246. Promptriever: Instruction-Trained Retrievers Can Be Prompted Like Language Models

链接：https://iclr.cc/virtual/2025/poster/28345 abstract： Instruction-tuned language models (LM) are able to respond to imperative commands, providing a more natural user interface compared to their base counterparts. In this work, we present Promptriever, the first retrieval model able to be prompted like an LM. To train Promptriever, we curate and release a new instance-level instruction training set from MS MARCO, spanning nearly 500k instances. Promptriever not only achieves strong performance on standard retrieval tasks, but also follows instructions. We observe: (1) large gains (reaching SoTA) on following detailed relevance instructions (+14.3 p-MRR / +3.1 nDCG on FollowIR), (2) significantly increased robustness to lexical choices/phrasing in the query+instruction (+12.9 Robustness@10 on InstructIR), and (3) the ability to perform hyper-parameter search via prompting to reliably improve retrieval performance (+1.4 average increase on BEIR). Promptriever demonstrates that retrieval models can be controlled with prompts on a per-query basis, setting the stage for future work aligning LM prompting techniques with information retrieval.

1247. AnyTouch: Learning Unified Static-Dynamic Representation across Multiple Visuo-tactile Sensors

链接：https://iclr.cc/virtual/2025/poster/29304 abstract： Visuo-tactile sensors aim to emulate human tactile perception, enabling robots to precisely understand and manipulate objects. Over time, numerous meticulously designed visuo-tactile sensors have been integrated into robotic systems, aiding in completing various tasks. However, the distinct data characteristics of these low-standardized visuo-tactile sensors hinder the establishment of a powerful tactile perception system. We consider that the key to addressing this issue lies in learning unified multi-sensor representations, thereby integrating the sensors and promoting tactile knowledge transfer between them. To achieve unified representation of this nature, we introduce TacQuad, an aligned multi-modal multi-sensor tactile dataset from four different visuo-tactile sensors, which enables the explicit integration of various sensors. Recognizing that humans perceive the physical environment by acquiring diverse tactile information such as texture and pressure changes, we further propose to learn unified multi-sensor representations from both static and dynamic perspectives. By integrating tactile images and videos, we present AnyTouch, a unified static-dynamic multi-sensor representation learning framework with a multi-level structure, aimed at both enhancing comprehensive perceptual abilities and enabling effective cross-sensor transfer. This multi-level architecture captures pixel-level details from tactile data via masked modeling and enhances perception and transferability by learning semantic-level sensor-agnostic features through multi-modal alignment and cross-sensor matching. We provide a comprehensive analysis of multi-sensor transferability, and validate our method on various offline datasets and in the real-world pouring task. Experimental results show that our method outperforms existing methods, exhibits outstanding static and dynamic perception capabilities across various sensors. The code, TacQuad dataset and AnyTouch model are fully available at gewu-lab.github.io/AnyTouch/.

1248. SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement

链接：https://iclr.cc/virtual/2025/poster/30299 abstract： Software engineers operating in complex and dynamic environments must continuously adapt to evolving requirements, learn iteratively from experience, and reconsider their approaches based on new insights. However, current large language model (LLM)-based software agents often follow linear, sequential processes that prevent backtracking and exploration of alternative solutions, limiting their ability to rethink their strategies when initial approaches prove ineffective. To address these challenges, we propose SWE-Search, a multi-agent framework that integrates Monte Carlo Tree Search (MCTS) with a self-improvement mechanism to enhance software agents' performance on repository-level software tasks. SWE-Search extends traditional MCTS by incorporating a hybrid value function that leverages LLMs for both numerical value estimation and qualitative evaluation. This enables self-feedback loops where agents iteratively refine their strategies based on both quantitative numerical evaluations and qualitative natural language assessments of pursued trajectories. The framework includes a SWE-Agent for adaptive exploration, a Value Agent for iterative feedback, and a Discriminator Agent that facilitates multi-agent debate for collaborative decision-making. Applied to the SWE-bench benchmark, our approach demonstrates a 23% relative improvement in performance across five models compared to standard open-source agents without MCTS. Our analysis reveals how performance scales with increased inference-time compute through deeper search, providing a pathway to improve software agents without requiring larger models or additional training data. This highlights the potential of self-evaluation driven search techniques in complex software engineering environments.

1249. Handling Delay in Real-Time Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29264 abstract： Real-time reinforcement learning (RL) introduces several challenges. First, policies are constrained to a fixed number of actions per second due to hardware limitations. Second, the environment may change while the network is still computing an action, leading to observational delay. The first issue can partly be addressed with pipelining, leading to higher throughput and potentially better policies. However, the second issue remains: if each neuron operates in parallel with an execution time of τ, an N-layer feed-forward network experiences observation delay of τN.Reducing the number of layers can decrease this delay, but at the cost of the network's expressivity. In this work, we explore the trade-off between minimizing delay and network's expressivity. We present a theoretically motivated solution that leverages temporal skip connections combined with history-augmented observations. We evaluate several architectures and show that those incorporating temporal skip connections achieve strong performance across various neuron execution times, reinforcement learning algorithms, and environments, including four Mujoco tasks and all MinAtar games. Moreover, we demonstrate parallel neuron computation can accelerate inference by 6-350\% on standard hardware. Our investigation into temporal skip connections and parallel computations paves the way for more efficient RL agents in real-time setting.

1250. Watermark Anything With Localized Messages

链接：https://iclr.cc/virtual/2025/poster/30152 abstract： Image watermarking methods are not tailored to handle small watermarked areas.This restricts applications in real-world scenarios where parts of the image may come from different sources or have been edited.We introduce a deep-learning model for localized image watermarking, dubbed the Watermark Anything Model (WAM). The WAM embedder imperceptibly modifies the input image, while the extractor segments the received image into watermarked and non-watermarked areas and recovers one or several hidden messages from the areas found to be watermarked.The models are jointly trained at low resolution and without perceptual constraints, then post-trained for imperceptibility and multiple watermarks.Experiments show that WAM is competitive with state-of-the art methods in terms of imperceptibility and robustness, especially against inpainting and splicing, even on high-resolution images. Moreover, it offers new capabilities: WAM can locate watermarked areas in spliced images and extract distinct 32-bit messages with less than 1 bit error from multiple small regions -- no larger than 10\% of the image surface -- even for small 256×256 images.Training and inference code and model weights are available at https://github.com/facebookresearch/watermark-anything.

1251. How Low Can You Go? Searching for the Intrinsic Dimensionality of Complex Networks using Metric Node Embeddings

链接：https://iclr.cc/virtual/2025/poster/29429 abstract： Low-dimensional embeddings are essential for machine learning tasks involving graphs, such as node classification, link prediction, community detection, network visualization, and network compression. Although recent studies have identified exact low-dimensional embeddings, the limits of the required embedding dimensions remain unclear. We presently prove that lower dimensional embeddings are possible when using Euclidean metric embeddings as opposed to vector-based Logistic PCA (LPCA) embeddings. In particular, we provide an efficient logarithmic search procedure for identifying the exact embedding dimension and demonstrate how metric embeddings enable inference of the exact embedding dimensions of large-scale networks by exploiting that the metric properties can be used to provide linearithmic scaling. Empirically, we show that our approach extracts substantially lower dimensional representations of networks than previously reported for small-sized networks. For the first time, we demonstrate that even large-scale networks can be effectively embedded in very low-dimensional spaces, and provide examples of scalable, exact reconstruction for graphs with up to a million nodes. Our approach highlights that the intrinsic dimensionality of networks is substantially lower than previously reported and provides a computationally efficient assessment of the exact embedding dimension also of large-scale networks. The surprisingly low dimensional representations achieved demonstrate that networks in general can be losslessly represented using very low dimensional feature spaces, which can be used to guide existing network analysis tasks from community detection and node classification to structure revealing exact network visualizations.

1252. CircuitFusion: Multimodal Circuit Representation Learning for Agile Chip Design

链接：https://iclr.cc/virtual/2025/poster/28186 abstract： The rapid advancements of AI rely on the support of integrated circuits (ICs). However, the growing complexity of digital ICs makes the traditional IC design process costly and time-consuming. In recent years, AI-assisted IC design methods have demonstrated great potential, but most methods are task-specific or focus solely on the circuit structure in graph format, overlooking other circuit modalities with rich functional information. In this paper, we introduce CircuitFusion, the first multimodal and implementation-aware circuit encoder. It encodes circuits into general representations that support different downstream circuit design tasks. To learn from circuits, we propose to fuse three circuit modalities: hardware code, structural graph, and functionality summary. More importantly, we identify four unique properties of circuits: parallel execution, functional equivalent transformation, multiple design stages, and circuit reusability. Based on these properties, we propose new strategies for both the development and application of CircuitFusion: 1) During circuit preprocessing, utilizing the parallel nature of circuits, we split each circuit into multiple sub-circuits based on sequential-element boundaries, each sub-circuit in three modalities. It enables fine-grained encoding at the sub-circuit level. 2) During CircuitFusion pre-training, we introduce three self-supervised tasks that utilize equivalent transformations both within and across modalities. We further utilize the multi-stage property of circuits to align representation with ultimate circuit implementation. 3) When applying CircuitFusion to downstream tasks, we propose a new retrieval-augmented inference method, which retrieves similar known circuits as a reference for predictions. It improves fine-tuning performance and even enables zero-shot inference. Evaluated on five different circuit design tasks, CircuitFusion consistently outperforms the state-of-the-art supervised method specifically developed for every single task, demonstrating its generalizability and ability to learn circuits' inherent properties.

1253. Can Generative AI Solve Your In-Context Learning Problem? A Martingale Perspective

链接：https://iclr.cc/virtual/2025/poster/29101 abstract： This work is about estimating when a conditional generative model (CGM) can solve an in-context learning (ICL) problem. An in-context learning (ICL) problem comprises a CGM, a dataset, and a prediction task. The CGM could be a multi-modal foundation model; the dataset, a collection of patient histories, test results, and recorded diagnoses; and the prediction task to communicate a diagnosis to a new patient. A Bayesian interpretation of ICL assumes that the CGM computes a posterior predictive distribution over an unknown Bayesian model defining a joint distribution over latent explanations and observable data. From this perspective, Bayesian model criticism is a reasonable approach to assess the suitability of a given CGM for an ICL problem. However, such approaches---like posterior predictive checks (PPCs)---often assume that we can sample from the likelihood and posterior defined by the Bayesian model, which are not explicitly given for contemporary CGMs. To address this, we show when ancestral sampling from the predictive distribution of a CGM is equivalent to sampling datasets from the posterior predictive of the assumed Bayesian model. Then we develop the generative predictive p-value, which enables PPCs and their cousins for contemporary CGMs. The generative predictive p-value can be used in a statistical decision procedure to determine when the model is appropriate for an ICL problem. Our method only requires generating queries and responses from a CGM and evaluating its response log probability. Using large language models, we empirically evaluate our method on tasks involving tabular data, imaging data, and natural language data.

1254. Lift Your Molecules: Molecular Graph Generation in Latent Euclidean Space

链接：https://iclr.cc/virtual/2025/poster/27981 abstract： We introduce a new framework for 2D molecular graph generation using 3D molecule generative models. Our Synthetic Coordinate Embedding (SyCo) framework maps 2D molecular graphs to 3D Euclidean point clouds via synthetic coordinates and learns the inverse map using an E(n)-Equivariant Graph Neural Network (EGNN). The induced point cloud-structured latent space is well-suited to apply existing 3D molecule generative models. This approach simplifies the graph generation problem into a point cloud generation problem followed by node and edge classification tasks, without relying on molecular fragments nor autoregressive decoding. Further, we propose a novel similarity-constrained optimization scheme for 3D diffusion models based on inpainting and guidance. As a concrete implementation of our framework, we develop EDM-SyCo based on the E(3) Equivariant Diffusion Model (EDM). EDM-SyCo achieves state-of-the-art performance in distribution learning of molecular graphs, outperforming the best non-autoregressive methods by more than 26\% on ZINC250K and 16\% on the GuacaMol dataset while improving conditional generation by up to 3.9 times.

1255. WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct

链接：https://iclr.cc/virtual/2025/poster/28471 abstract： Large language models (LLMs), such as GPT-4, have shown remarkable performance in natural language processing (NLP) tasks, including challenging mathematical reasoning. However, most existing open-source models are only pre-trained on large-scale internet data and without math-related optimization. In this paper, we present WizardMath, which enhances the mathematical reasoning abilities of LLMs, by applying our proposed Reinforcement Learning from Evol-Instruct Feedback (RLEIF) method to the domain of math. Through extensive experiments on two mathematical reasoning benchmarks, namely GSM8k and MATH, we reveal the extraordinary capabilities of our model. Remarkably, WizardMath-Mistral 7B surpasses all other open-source LLMs by a substantial margin. Furthermore, WizardMath 70B even outperforms ChatGPT-3.5, Claude Instant, Gemini Pro and Mistral Medium. Additionally, our preliminary exploration highlights the pivotal role of instruction evolution and process supervision in achieving exceptional math performance.

1256. Statistical Tractability of Off-policy Evaluation of History-dependent Policies in POMDPs

链接：https://iclr.cc/virtual/2025/poster/29675 abstract： We investigate off-policy evaluation (OPE), a central and fundamental problemin reinforcement learning (RL), in the challenging setting of Partially ObservableMarkov Decision Processes (POMDPs) with large observation spaces. Recentworks of Uehara et al. (2023a); Zhang & Jiang (2024) developed a model-freeframework and identified important coverage assumptions (called belief and outcome coverage) that enable accurate OPE of memoryless policies with polynomial sample complexities, but handling more general target policies that depend onthe entire observable history remained an open problem. In this work, we proveinformation-theoretic hardness for model-free OPE of history-dependent policies inseveral settings, characterized by additional assumptions imposed on the behaviorpolicy (memoryless vs. history-dependent) and/or the state-revealing property ofthe POMDP (single-step vs. multi-step revealing). We further show that some hardness can be circumvented by a natural model-based algorithm—whose analysis has surprisingly eluded the literature despite the algorithm’s simplicity—demonstratingprovable separation between model-free and model-based OPE in POMDPs.

1257. Iterative Nash Policy Optimization: Aligning LLMs with General Preferences via No-Regret Learning

链接：https://iclr.cc/virtual/2025/poster/29724 abstract： Reinforcement Learning with Human Feedback (RLHF) has achieved great successin aligning large language models (LLMs) with human preferences. PrevalentRLHF approaches are reward-based, following the Bradley-Terry (BT) model assumption, which may not fully capture the complexity of human preferences. Inthis paper, we explore RLHF under a general preference framework and approachit from a game-theoretic perspective. Specifically, we formulate the problem asa two-player game and propose a novel online algorithm, iterative Nash policyoptimization (INPO). The key idea is to let the policy play against itself via no-regret learning, thereby approximating the Nash policy. Unlike previous methods,INPO bypasses the need for estimating the expected win rate for individual responses, which typically incurs high computational or annotation costs. Instead,we introduce a new loss objective that is directly minimized over a preferencedataset. We provide theoretical analysis for our approach and demonstrate itseffectiveness through experiments on various representative benchmarks. With anLLaMA-3-8B-based SFT model, INPO achieves a 42.6% length-controlled winrate on AlpacaEval 2.0 and a 37.8% win rate on Arena-Hard, showing substantialimprovement over the state-of-the-art online RLHF algorithms.

1258. SPA: 3D Spatial-Awareness Enables Effective Embodied Representation

链接：https://iclr.cc/virtual/2025/poster/30883 abstract： In this paper, we introduce SPA, a novel representation learning framework that emphasizes the importance of 3D spatial awareness in embodied AI. Our approach leverages differentiable neural rendering on multi-view images to endow a vanilla Vision Transformer (ViT) with intrinsic spatial understanding. We present the most comprehensive evaluation of embodied representation learning to date, covering 268 tasks across 8 simulators with diverse policies in both single-task and language-conditioned multi-task scenarios. The results are compelling: SPA consistently outperforms more than 10 state-of-the-art representation methods, including those specifically designed for embodied AI, vision-centric tasks, and multi-modal applications, while using less training data. Furthermore, we conduct a series of real-world experiments to confirm its effectiveness in practical scenarios. These results highlight the critical role of 3D spatial awareness for embodied representation learning. Our strongest model takes more than 6000 GPU hours to train and we are committed to open-sourcing all code and model weights to foster future research in embodied representation learning.

1259. COAT: Compressing Optimizer states and Activations for Memory-

Efficient FP8 Training

链接：https://iclr.cc/virtual/2025/poster/29297 abstract： FP8 training has emerged as a promising method for improving training efficiency. Existing frameworks accelerate training by applying FP8 computation to linear layers while leaving optimizer states and activations in higher precision, which fails to fully optimize memory usage. This paper introduces COAT (Compressing Optimizer States and Activations for FP8 Training), a novel FP8 training framework designed to significantly reduce memory footprint when training large models. COAT addresses current limitations through two key innovations: (1) Dynamic Range Expansion, which aligns optimizer state distributions more closely with the FP8 representation range, thereby reducing quantization error, and (2) Mixed-Granularity Activation Quantization, which optimizes activation memory using a combination of per-tensor and per-group quantization strategies. Experiments demonstrate that COAT effectively reduces end-to-end training memory footprint by 1.54× compared to BF16 while achieving nearly lossless performance across various tasks, such as Large Language Model pretraining and fine-tuning and Vision Language Model training. COAT also achieves a 1.43× end-to-end training speedup compared to BF16, performing on par with or surpassing TransformerEngine's speedup. COAT enables efficient full-parameter training of large models on fewer GPUs, and facilitates doubling the batch size in distributed training settings, providing a practical solution for scaling large-scale model training. Code will be released upon publication.

1260. Mechanistic Interpretability Meets Vision Language Models: Insights and Limitations

链接：https://iclr.cc/virtual/2025/poster/31330 abstract： Vision language models (VLMs), such as GPT-4o, have rapidly evolved, demonstrating impressive capabilities across diverse tasks. However, much of the progress in this field has been driven by engineering efforts, with a limited understanding of how these models work. The lack of scientific insight poses challenges to further enhancing their robustness, generalization, and interpretability, especially in high-stakes settings. In this work, we systematically review the use of mechanistic interpretability methods to foster a more scientific and transparent understanding of VLMs. Specifically, we examine five prominent techniques: probing, activation patching, logit lens, sparse autoencoders, and automated explanation. We summarize the key insights these methods provide into how VLMs process information and make decisions. We also discuss critical challenges and limitations that must be addressed to further advance the field.

1261. A Little Goes a Long Way: Efficient Long Context Training and Inference with Partial Contexts

链接：https://iclr.cc/virtual/2025/poster/29510 abstract： Training and serving long-context large language models (LLMs) incurs substantial overhead. To address this, two critical steps are often required: a pretrained LLM typically undergoes a separate stage for context length extension by training on long-context data, followed by architectural modifications to reduce the overhead of KV cache during serving. This paper argues that integrating length extension with a GPU-friendly KV cache reduction architecture not only reduces training overhead during length extension, but also achieves better long-context performance. This leads to our proposed LongGen, which finetunes a pretrained LLM into an efficient architecture during length extension. LongGen builds on three key insights: (1) Sparse attention patterns, such as window attention (attending to recent tokens), attention sink (initial ones), and blockwise sparse attention (strided token blocks) are well-suited for building efficient long-context models, primarily due to their GPU-friendly memory access patterns, enabling efficiency gains not just theoretically but in practice as well. (2) It is essential for the model to have direct access to all tokens. A hybrid architecture with 1/3 full attention layers and 2/3 efficient ones achieves a balanced trade-off between efficiency and long-context performance.(3) Lightweight training on 5B long-context data is sufficient to extend the hybrid model's context length from 4K to 128K.We evaluate LongGen on both Llama-2 7B and Llama-2 70B, demonstrating its effectiveness across different scales. During training with 128K-long contexts, LongGen achieves 1.55x training speedup and reduces wall-clock time by 36%, compared to a full-attention baseline. During inference, LongGen reduces KV cache memory by 62%, achieving 1.67x prefilling speedup and 1.41x decoding speedup.Compared to baselines that apply KV-cache reduction techniques to full-attention long-context LLMs, LongGen achieves substantially stronger performance not only on the Needle-in-a-Haystack retrieval task, but also on more challenging long-context reasoning tasks, including BABILong and RULER.

1262. Efficient and Trustworthy Causal Discovery with Latent Variables and Complex Relations

链接：https://iclr.cc/virtual/2025/poster/30566 abstract： Most traditional causal discovery methods assume that all task-relevant variables are observed, an assumption often violated in practice. Although some recent works allow the presence of latent variables, they typically assume the absence of certain special causal relations to ensure a degree of simplicity, which might also be invalid in real-world scenarios. This paper tackles a challenging and important setting where latent and observed variables are interconnected through complex causal relations. Under a pure children assumption ensuring that latent variables leave adequate footprints in observed variables, we develop novel theoretical results, leading to an efficient causal discovery algorithm which is the first one capable of handling the setting with both latent variables and complex relations within polynomial time. Our algorithm first sequentially identifies latent variables from leaves to roots and then sequentially infers causal relations from roots to leaves. Moreover, we prove trustworthiness of our algorithm, meaning that when the assumption is invalid, it can raise an error signal rather than draw an incorrect causal conclusion, thus preventing potential damage to downstream tasks. We demonstrate the efficacy of our algorithm through experiments. Our work significantly enhances efficiency and reliability of causal discovery in complex systems. Our code is available at: https://github.com/XiuchuanLi/ICLR2025-ETCD

1263. Video Action Differencing

链接：https://iclr.cc/virtual/2025/poster/31067 abstract： How do two individuals differ when performing the same action? In this work, we introduce Video Action Differencing (VidDiff), the novel task of identifying subtle differences between videos of the same action, which has numerous applications, such as coaching and skill learning. To enable development on this new task, we first create VidDiffBench, a benchmark dataset containing 549 video pairs, with human annotations of 4,469 fine-grained action differences and 2,075 timestamps indicating where these differences occur. Our experiments demonstrate that VidDiffBench poses a significant challenge for state-of-the-art large multimodal models (LMMs), such as GPT-4o and Qwen2-VL. By analyzing the failure cases of LMMs on VidDiffBench, we highlight two key challenges for this task: localizing relevant sub-actions over two videos and fine-grained frame comparison. To overcome these, we propose the VidDiff method, an agentic workflow that breaks the task into three stages: action difference proposal, keyframe localization, and frame differencing, each stage utilizing specialized foundation models. To encourage future research in this new task, we release the benchmark and code.

1264. Hierarchical Uncertainty Estimation for Learning-based Registration in Neuroimaging

链接：https://iclr.cc/virtual/2025/poster/27869 abstract： Over recent years, deep learning based image registration has achieved impressive accuracy in many domains, including medical imaging and, specifically, human neuroimaging with magnetic resonance imaging (MRI). However, the uncertainty estimation associated with these methods has been largely limited to the application of generic techniques (e.g., Monte Carlo dropout) that do not exploit the peculiarities of the problem domain, particularly spatial modeling. Here, we propose a principled way to propagate uncertainties (epistemic or aleatoric) estimated at the level of spatial location by these methods, to the level of global transformation models, and further to downstream tasks. Specifically, we justify the choice of a Gaussian distribution for the local uncertainty modeling, and then propose a framework where uncertainties spread across hierarchical levels, depending on the choice of transformation model. Experiments on publicly available data sets show that Monte Carlo dropout correlates very poorly with the reference registration error, whereas our uncertainty estimates correlate much better. Crucially, the results also show that uncertainty-aware fitting of transformations improves the registration accuracy of brain MRI scans. Finally, we illustrate how sampling from the posterior distribution of the transformations can be used to propagate uncertainties to downstream neuroimaging tasks. Code is available at: https://github.com/HuXiaoling/Regre4Regis.

1265. Recovery of Causal Graph Involving Latent Variables via Homologous Surrogates

链接：https://iclr.cc/virtual/2025/poster/28885 abstract： Causal discovery with latent variables is an important and challenging problem. To identify latent variables and infer their causal relations, most existing works rely on the assumption that latent variables have pure children. Considering that this assumption is potentially restrictive in practice and not strictly necessary in theory, in this paper, by introducing the concept of homologous surrogate, we eliminate the need for pure children in the context of causal discovery with latent variables. The homologous surrogate fundamentally differs from the pure child in the sense that the latter is characterized by having strictly restricted parents while the former allows for much more flexible parents. We formulate two assumptions involving homologous surrogates and develop theoretical results under each assumption. Under the weaker assumption, our theoretical results imply that we can determine each variable's ancestors, that is, partially recover the causal graph. The stronger assumption further enables us to determine each variable's parents exactly, that is, fully recover the causal graph. Building on these theoretical results, we derive an algorithm that fully leverages the properties of homologous surrogates for causal graph recovery. Also, we validate its efficacy through experiments. Our work broadens the applicability of causal discovery. Our code is available at: https://github.com/XiuchuanLi/ICLR2025-CDHS

1266. GenEx: Generating an Explorable World

链接：https://iclr.cc/virtual/2025/poster/30770 abstract： Understanding, navigating, and exploring the 3D physical real world has long been a central challenge in the development of artificial intelligence. In this work, we take a step toward this goal by introducing GenEx, a system capable of planning complex embodied world exploration, guided by its generative imagination that forms expectations about the surrounding environments. GenEx generates high-quality, continuous 360-degree virtual environments, achieving robust loop consistency and active 3D mapping over extended trajectories. Leveraging generative imagination, GPT-assisted agents can undertake complex embodied tasks, including goal-agnostic exploration and goal-driven navigation. Agents utilize imagined observations to update their beliefs, simulate potential outcomes, and enhance their decision-making. Training on the synthetic urban dataset GenEx-DB and evaluation on GenEx-EQA demonstrate that our approach significantly improves agents' planning capabilities, providing a transformative platform toward intelligent, imaginative embodied exploration.

1267. Understanding and Enhancing the Transferability of Jailbreaking Attacks

链接：https://iclr.cc/virtual/2025/poster/29142 abstract： Jailbreaking attacks can effectively manipulate open-source large language models (LLMs) to produce harmful responses. However, these attacks exhibit limited transferability, failing to disrupt proprietary LLMs consistently. To reliably identify vulnerabilities in proprietary LLMs, this work investigates the transferability of jailbreaking attacks by analysing their impact on the model's intent perception. By incorporating adversarial sequences, these attacks can redirect the source LLM's focus away from malicious-intent tokens in the original input, thereby obstructing the model's intent recognition and eliciting harmful responses. Nevertheless, these adversarial sequences fail to mislead the target LLM's intent perception, allowing the target LLM to refocus on malicious-intent tokens and abstain from responding. Our analysis further reveals the inherent $\textit{distributional dependency}$ within the generated adversarial sequences, whose effectiveness stems from overfitting the source LLM's parameters, resulting in limited transferability to target LLMs. To this end, we propose the Perceived-importance Flatten (PiF) method, which uniformly disperses the model's focus across neutral-intent tokens in the original input, thus obscuring malicious-intent tokens without relying on overfitted adversarial sequences. Extensive experiments demonstrate that PiF provides an effective and efficient red-teaming evaluation for proprietary LLMs.

1268. Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering

链接：https://iclr.cc/virtual/2025/poster/30879 abstract： For vision-language models (VLMs), understanding the dynamic properties of objects and their interactions in 3D scenes from videos is crucial for effective reasoning about high-level temporal and action semantics. Although humans are adept at understanding these properties by constructing 3D and temporal (4D) representations of the world, current video understanding models struggle to extract these dynamic semantics, arguably because these models use cross-frame reasoning without underlying knowledge of the 3D/4D scenes.In this work, we introduce DynSuperCLEVR, the first video question answering dataset that focuses on language understanding of the dynamic properties of 3D objects. We concentrate on three physical concepts—velocity, acceleration, and collisions—within 4D scenes. We further generate three types of questions, including factual queries, future predictions, and counterfactual reasoning that involve different aspects of reasoning on these 4D dynamic properties.To further demonstrate the importance of explicit scene representations in answering these 4D dynamics questions, we propose NS-4DPhysics, a Neural-Symbolic VideoQA model integrating Physics prior for 4D dynamic properties with explicit scene representation of videos. Instead of answering the questions directly from the video text input, our method first estimates the 4D world states with a 3D generative model powered by a physical prior, and then uses neural symbolic reasoning to answer the questions based on the 4D world states.Our evaluation on all three types of questions in DynSuperCLEVR shows that previous video question answering models and large multimodal models struggle with questions about 4D dynamics, while our NS-4DPhysics significantly outperforms previous state-of-the-art models.

1269. Training LLMs over Neurally Compressed Text

链接：https://iclr.cc/virtual/2025/poster/31457 abstract： In this paper, we explore the idea of training large language models (LLMs) over highly compressed text. While standard subword tokenizers compress text by a small factor, neural text compressors can achieve much higher rates of compression. If it were possible to train LLMs directly over neurally compressed text, this would confer advantages in training and serving efficiency, as well as easier handling of long text spans. The main obstacle to this goal is that strong compression tends to produce opaque outputs that are not well-suited for learning. In particular, we find that text naïvely compressed via Arithmetic Coding is not readily learnable by LLMs. To overcome this, we propose Equal-Info Windows, a novel compression technique whereby text is segmented into blocks that each compress to the same bit length. Using this method, we demonstrate effective learning over neurally compressed text that improves with scale, and outperforms byte-level baselines by a wide margin on perplexity and inference speed benchmarks. While our method delivers worse perplexity than subword tokenizers for models trained with the same parameter count, it has the benefit of shorter sequence lengths. Shorter sequence lengths require fewer autoregressive generation steps, often reducing latency. Finally, we provide extensive analysis of the properties that contribute to learnability, and offer concrete suggestions for how to further improve the performance of high-compression tokenizers.

1270. Autoregressive Pretraining with Mamba in Vision

链接：https://iclr.cc/virtual/2025/poster/29755 abstract： The vision community has started to build with the recently developed state space model, Mamba, as the new backbone for a range of tasks. This paper shows that Mamba's visual capability can be significantly enhanced through autoregressive pretraining, a direction not previously explored. Efficiency-wise, the autoregressive nature can well capitalize on the Mamba's unidirectional recurrent structure, enabling faster overall training speed compared to other training strategies like mask modeling. Performance-wise, autoregressive pretraining equips the Mamba architecture with markedly higher accuracy over its supervised-trained counterparts and, more importantly, successfully unlocks its scaling potential to large and even huge model sizes. For example, with autoregressive pretraining, a base-size Mamba attains 83.2\% ImageNet accuracy, outperforming its supervised counterpart by 2.0\%; our huge-size Mamba, the largest Vision Mamba to date, attains 85.0\% ImageNet accuracy (85.5\% when finetuned with 384×384 inputs), notably surpassing all other Mamba variants in vision. The code is available at \url{https://github.com/OliverRensu/ARM}.

1271. Mastering Task Arithmetic: τJp as a Key Indicator for Weight Disentanglement

链接：https://iclr.cc/virtual/2025/poster/31198 abstract： Model-editing techniques using task arithmetic have rapidly gained attention.Through task arithmetic, simply through arithmetic operations on the weights of pre-trained and fine-tuned models create desired models, such as multi-task models, models in which specific tasks are unsolvable, or domain-transferred models.However, task arithmetic faces challenges, such as poor reproducibility and the high cost associated with adjusting coefficients in the arithmetic operations on model parameters, which have limited its practical success. In this paper, we present three key contributions in the context of task addition and task negation within task arithmetic. First, we propose a new metric called τJp which is based on the product of the task vector (τ) and the Jacobian of the pre-trained model with respect to its weights. We show that τJp has a causal relationship with the interference that occurs from arithmetic operations. Second, we show that introducing regularization to minimize τJp significantly mitigates interference between task inference, which leads to the elimination of coefficient tuning and improved accuracy on each task.Third, in the context of incremental learning, we demonstrate that our τJp regularization achieves more robust performance in environments where access to future tasks is unavailable, thus validating the scalability of the approach.Finally, we demonstrate that the τJp regularizer further reinforces the performance of task arithmetic by leveraging publicly available fine-tuned models, offering practical benefits for real-world applications.Our code is available at https://github.com/katoro8989/tau-Jp_Task_Arithmetic

1272. Instance-dependent Early Stopping

链接：https://iclr.cc/virtual/2025/poster/29782 abstract： In machine learning practice, early stopping has been widely used to regularize models and can save computational costs by halting the training process when the model's performance on a validation set stops improving. However, conventional early stopping applies the same stopping criterion to all instances without considering their individual learning statuses, which leads to redundant computations on instances that are already well-learned. To further improve the efficiency, we propose an Instance-dependent Early Stopping (IES) method that adapts the early stopping mechanism from the entire training set to the instance level, based on the core principle that once the model has mastered an instance, the training on it should stop. IES considers an instance as mastered if the second-order differences of its loss value remain within a small range around zero. This offers a more consistent measure of an instance's learning status compared with directly using the loss value, and thus allows for a unified threshold to determine when an instance can be excluded from further backpropagation. We show that excluding mastered instances from backpropagation can increase the gradient norms, thereby accelerating the decrease of the training loss and speeding up the training process. Extensive experiments on benchmarks demonstrate that IES method can reduce backpropagation instances by 10%-50% while maintaining or even slightly improving the test accuracy and transfer learning performance of a model.

1273. SFS: Smarter Code Space Search improves LLM Inference Scaling

链接：https://iclr.cc/virtual/2025/poster/29947 abstract： We frame code generation as a black-box optimization problem within the codespace and demonstrate how optimization-inspired techniques can enhance inferencescaling over text. Based on this perspective, we propose SCATTERED FORESTSEARCH (SFS), a novel approach that improves solution diversity during evolutionary search,thereby avoiding local optima. Our theoretical analysis illustrates how thesemethods improve exploration and enhance efficiency. Extensive experimentson HumanEval, MBPP, APPS, CodeContests, and Leetcode reveal significantperformance gains. For instance, our method achieves a pass@1 rate of 67.1% onHumanEval+ and 87.2% on HumanEval with GPT-3.5, marking improvements of8.6% and 4.3% over the state-of-the-art, while also halving the iterations neededto find the correct solution. Furthermore, our approach scales more efficientlythan existing search techniques, including tree search, line search, and repeatedsampling (Best of N).

1274. YOLO-RD: Introducing Relevant and Compact Explicit Knowledge to YOLO by Retriever-Dictionary

链接：https://iclr.cc/virtual/2025/poster/30052 abstract： Identifying and localizing objects within images is a fundamental challenge, and numerous efforts have been made to enhance model accuracy by experimenting with diverse architectures and refining training strategies. Nevertheless, a prevalent limitation in existing models is overemphasizing the current input while ignoring the information from the entire dataset. We introduce an innovative $\textbf{R}etriever-\textbf{D}ictionary$ (RD) module to address this issue. This architecture enables YOLO-based models to efficiently retrieve features from a Dictionary that contains the insight of the dataset, which is built by the knowledge from Visual Models (VM), Large Language Models (LLM), or Visual Language Models (VLM). The flexible RD enables the model to incorporate such explicit knowledge that enhances the ability to benefit multiple tasks, specifically, segmentation, detection, and classification, from pixel to image level. The experiments show that using the RD significantly improves model performance, achieving more than a 3\% increase in mean Average Precision for object detection with less than a 1\% increase in model parameters. Beyond 1-stage object detection models, the RD module improves the effectiveness of 2-stage models and DETR-based architectures, such as Faster R-CNN and Deformable DETR. Code is released at https://github.com/henrytsui000/YOLO.

1275. Fictitious Synthetic Data Can Improve LLM Factuality via Prerequisite Learning

链接：https://iclr.cc/virtual/2025/poster/29436 abstract： Recent studies have identified one aggravating factor of LLM hallucinations as the knowledge inconsistency between pre-training and fine-tuning, where unfamiliar fine-tuning data mislead the LLM to fabricate plausible but wrong outputs. In this paper, we propose a novel fine-tuning strategy called Prereq-Tune to address this knowledge inconsistency and reduce hallucinations. Fundamentally, Prereq-Tune disentangles the learning of skills and knowledge, so the model learns only the task skills without being impacted by the knowledge inconsistency. To achieve this, Prereq-Tune introduces an additional prerequisite learning stage to learn the necessary knowledge for SFT, allowing subsequent SFT to focus only on task skills. Prereq-Tune can also be combined with fictitious synthetic data to enhance the grounding of LLM outputs to their internal knowledge. Experiments show that Prereq-Tune outperforms existing baselines in improving LLM's factuality across short QA and long-form generation tasks. It also opens new possibilities for knowledge-controlled generation in LLMs. Our code is available at https://github.com/UCSB-NLP-Chang/Prereq_tune.git.

1276. Rational Decision-Making Agent with Learning Internal Utility Judgment

链接：https://iclr.cc/virtual/2025/poster/30292 abstract： With remarkable advancements, large language models (LLMs) have attracted significant efforts to develop LLM-based agents capable of executing intricate multi-step decision-making tasks. Existing approaches predominantly build upon the external performance measure to guide the decision-making process but the reliance on the external performance measure as prior is problematic in real-world scenarios, where such prior may be unavailable, flawed, or even erroneous. For genuine autonomous decision-making for LLM-based agents, it is imperative to develop rationality from their posterior experiences to judge the utility of each decision independently. In this work, we propose RaDAgent (Rational Decision-Making Agent), which fosters the development of its rationality through an iterative framework involving Experience Exploration and Utility Learning. Within this framework, Elo-based Utility Learning is devised to assign Elo scores to individual decision steps to judge their utilities via pairwise comparisons. Consequently, these Elo scores guide the decision-making process to derive optimal outcomes. Experimental results on the Game of 24, WebShop, ToolBench and RestBench datasets demonstrate RaDAgent’s superiority over baselines, achieving about 7.8% improvement on average. Besides, RaDAgent also can reduce costs (ChatGPT API calls), highlighting its effectiveness and efficiency.

1277. IFORMER: INTEGRATING CONVNET AND TRANSFORMER FOR MOBILE APPLICATION

链接：https://iclr.cc/virtual/2025/poster/30980 abstract： We present a new family of mobile hybrid vision networks, called iFormer, with a focus on optimizing latency and accuracy on mobile applications. iFormer effectively integrates the fast local representation capacity of convolution with the efficient global modeling ability of self-attention. The local interactions are derived from transforming a standard convolutional network, \textit{i.e.}, ConvNeXt, to design a more lightweight mobile network. Our newly introduced mobile modulation attention removes memory-intensive operations in MHA and employs an efficient modulation mechanism to boost dynamic global representational capacity. We conduct comprehensive experiments demonstrating that iFormer outperforms existing lightweight networks across various tasks. Notably, iFormer achieves an impressive Top-1 accuracy of 80.4% on ImageNet-1k with a latency of only 1.10 ms on an iPhone 13, surpassing the recently proposed MobileNetV4 under similar latency constraints. Additionally, our method shows significant improvements in downstream tasks, including COCO object detection, instance segmentation, and ADE20k semantic segmentation, while still maintaining low latency on mobile devices for high-resolution inputs in these scenarios. Code and models are available at: https://github.com/ChuanyangZheng/iFormer.

1278. TypedThinker: Diversify Large Language Model Reasoning with Typed Thinking

链接：https://iclr.cc/virtual/2025/poster/29421 abstract： Large Language Models (LLMs) have demonstrated strong reasoning capabilities in solving complex problems. However, current approaches primarily enhance reasoning through the elaboration of thoughts while neglecting the diversity of reasoning types. LLMs typically employ deductive reasoning, proceeding step-by-step from given conditions, which limits their exploration during problem-solving. Our analysis reveals that certain problems are exclusively solvable through specific reasoning strategies like inductive, abductive, or analogical reasoning. However, incorporating diverse reasoning approaches presents two key challenges: identifying the appropriate reasoning type for each problem and exploiting this approach during problem-solving. Therefore, we propose the TypedThinker that predicts suitable reasoning types based on the problem and their previous effectiveness and provides relevant demonstrations to guide LLMs in applying these strategies. Experimental results show significant improvements across multiple benchmarks, with performance gains of 3.4\% for Mistral 7B, 6.5\% for LLaMA3 8B, and 7\% for Qwen 2 7B on logical and mathematical reasoning tasks. TypedThinker enhances LLM reasoning without requiring knowledge distillation from larger models. It can be integrated into more advanced systems like GPT-4o or specialized models like MetaMath to diversify their reasoning approaches and improve their problem-solving capabilities.

1279. Identification of Intermittent Temporal Latent Process

链接：https://iclr.cc/virtual/2025/poster/30889 abstract： Identifying time-delayed latent causal process is crucial for understanding temporal dynamics and enabling downstream reasoning. While recent methods have made progress in identifying latent time-delayed causal processes, they cannot address the dynamics in which the influence of some latent factors on both the subsequent latent states and the observed data can become inactive or irrelevant at different time steps. Therefore, we introduce intermittent temporal latent processes, where: (1) any subset of latent factors may be missing during nonlinear data generation at any time step, and (2) the active latent factors at each step are unknown. This framework encompasses both nonstationary and stationary transitions, accommodating changing or consistent active factors over time. Our work shows that under certain assumptions, the latent causal variables are block-wise identifiable. With further conditional independence assumption, each latent variable can even be recovered up to component-wise transformations. Using this identification theory, we propose an unsupervised approach, InterLatent, to reliably uncover the representations of the intermittent temporal latent process. The experimental findings on both synthetic and real-world datasets verify our theoretical claims.

1280. Failures to Find Transferable Image Jailbreaks Between Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/27813 abstract： The integration of new modalities into frontier AI systems offers exciting capabilities, but also increases the possibility such systems can be adversarially manipulated in undesirable ways.In this work, we focus on a popular class of vision-language models (VLMs) that generate text outputs conditioned on visual and textual inputs.We conducted a large-scale empirical study to assess the transferability of gradient-based universal image "jailbreaks" using a diverse set of over 40 open-parameter VLMs, including 18 new VLMs that we publicly release.Overall, we find that transferable gradient-based image jailbreaks are extremely difficult to obtain.When an image jailbreak is optimized against a single VLM or against an ensemble of VLMs, the jailbreak successfully jailbreaks the attacked VLM(s), but exhibits little-to-no transfer to any other VLMs; transfer is not affected by whether the attacked and target VLMs possess matching vision backbones or language models, whether the language model underwent instruction-following and/or safety-alignment training, or many other factors.Only two settings display partially successful transfer: between identically-pretrained and identically-initialized VLMs with slightly different VLM training data, and between different training checkpoints of a single VLM.Leveraging these results, we then demonstrate that transfer can be significantly improved against a specific target VLM by attacking larger ensembles of "highly-similar" VLMs.These results stand in stark contrast to existing evidence of universal and transferable text jailbreaks against language models and transferable adversarial attacks against image classifiers, suggesting that VLMs may be more robust to gradient-based transfer attacks.

1281. Synergy Between Sufficient Changes and Sparse Mixing Procedure for Disentangled Representation Learning

链接：https://iclr.cc/virtual/2025/poster/30309 abstract： Disentangled representation learning aims to uncover the latent variables underlying observed data, yet identifying these variables under mild assumptions remains challenging. Some methods rely on sufficient changes in the distribution of latent variables indicated by auxiliary variables, such as domain indices, but acquiring enough domains is often impractical. Alternative approaches exploit the structural sparsity assumption on mixing processes, but this constraint may not hold in practice. Interestingly, we find that these two seemingly unrelated assumptions can actually complement each other. Specifically, when conditioned on auxiliary variables, the sparse mixing process induces independence between latent and observed variables, which simplifies the mapping from estimated to true latent variables and hence compensates for deficiencies of auxiliary variables. Building on this insight, we propose an identifiability theory with less restrictive constraints regarding the auxiliary variables and the sparse mixing process, enhancing applicability to real-world scenarios. Additionally, we develop a generative model framework incorporating a domain encoding network and a sparse mixing constraint and provide two implementations based on variational autoencoders and generative adversarial networks. Experiment results on synthetic and real-world datasets support our theoretical results.

1282. Causal Representation Learning from Multimodal Biomedical Observations

链接：https://iclr.cc/virtual/2025/poster/28747 abstract： Prevalent in biomedical applications (e.g., human phenotype research), multimodal datasets can provide valuable insights into the underlying physiological mechanisms. However, current machine learning (ML) models designed to analyze these datasets often lack interpretability and identifiability guarantees, which are essential for biomedical research. Recent advances in causal representation learning have shown promise in identifying interpretable latent causal variables with formal theoretical guarantees. Unfortunately, most current work on multimodal distributions either relies on restrictive parametric assumptions or yields only coarse identification results, limiting their applicability to biomedical research that favors a detailed understanding of the mechanisms.In this work, we aim to develop flexible identification conditions for multimodal data and principled methods to facilitate the understanding of biomedical datasets. Theoretically, we consider a nonparametric latent distribution (c.f., parametric assumptions in previous work) that allows for causal relationships across potentially different modalities. We establish identifiability guarantees for each latent component, extending the subspace identification results from previous work. Our key theoretical contribution is the structural sparsity of causal connections between modalities, which, as we will discuss, is natural for a large collection of biomedical systems.Empirically, we present a practical framework to instantiate our theoretical insights. We demonstrate the effectiveness of our approach through extensive experiments on both numerical and synthetic datasets. Results on a real-world human phenotype dataset are consistent with established biomedical research, validating our theoretical and methodological framework.

1283. Diffusion-NPO: Negative Preference Optimization for Better Preference Aligned Generation of Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/32065 abstract： Diffusion models have made substantial advances in image generation, yet models trained on large, unfiltered datasets often yield outputs misaligned with human preferences. Numerous methods have already been proposed to fine-tune pre-trained diffusion models, achieving notable improvements in aligning generated outputs with human preferences. However, we point out that existing preference alignment methods neglect the critical role of handling unconditional/negative-conditional outputs, leading to a diminished capacity to avoid generating undesirable outcomes. This oversight limits the efficacy of classifier-free guidance (CFG), which relies on the contrast between conditional generation and unconditional/negative-conditional generation to optimize output quality. In response, we propose a straightforward but consistently effective approach that involves training a model specifically attuned to negative preferences. This method does not require new training strategies or datasets but rather involves minor modifications to existing techniques. Our approach integrates seamlessly with models such as SD15, SDXL, video diffusion models and models that have undergone preference optimization, consistently enhancing their ability to produce more human preferences aligned outputs.

1284. Sparse autoencoders reveal selective remapping of visual concepts during adaptation

链接：https://iclr.cc/virtual/2025/poster/28675 abstract： Adapting foundation models for specific purposes has become a standard approach to build machine learning systems for downstream applications. Yet, it is an open question which mechanisms take place during adaptation. Here we develop a new Sparse Autoencoder (SAE) for the CLIP vision transformer, named PatchSAE, to extract interpretable concepts at granular levels (e.g., shape, color, or semantics of an object) and their patch-wise spatial attributions. We explore how these concepts influence the model output in downstream image classification tasks and investigate how recent state-of-the-art prompt-based adaptation techniques change the association of model inputs to these concepts. While activations of concepts slightly change between adapted and non-adapted models, we find that the majority of gains on common adaptation tasks can be explained with the existing concepts already present in the non-adapted foundation model. This work provides a concrete framework to train and use SAEs for Vision Transformers and provides insights into explaining adaptation mechanisms.

1285. How much of my dataset did you use? Quantitative Data Usage Inference in Machine Learning

链接：https://iclr.cc/virtual/2025/poster/30403 abstract： How much of my data was used to train a machine learning model? This is a critical question for data owners assessing the risk of unauthorized usage of their data to train models. However, previous work mistakenly treats this as a binary problem—inferring whether all-or-none or any-or-none of the data was used—which is fragile when faced with real, non-binary data usage risks. To address this, we propose a fine-grained analysis called Dataset Usage Cardinality Inference (DUCI), which estimates the exact proportion of data used. Our algorithm, leveraging debiased membership guesses, matches the performance of the optimal MLE approach (with a maximum error <0.1) but with significantly lower (e.g., $300 \times$ less) computational cost.

1286. Discovering Influential Neuron Path in Vision Transformers

链接：https://iclr.cc/virtual/2025/poster/29362 abstract： Vision Transformer models exhibit immense power yet remain opaque to human understanding, posing challenges and risks for practical applications. While prior research has attempted to demystify these models through input attribution and neuron role analysis,there's been a notable gap in considering layer-level information and the holistic path of information flow across layers.In this paper, we investigate the significance of influential neuron paths within vision Transformers, which is a path of neurons from the model input to output that impacts the model inference most significantly.We first propose a joint influence measure to assess the contribution of a set of neurons to the model outcome.And we further provide a layer-progressive neuron locatingapproach that efficiently selects the most influential neuron at each layer trying to discover the crucial neuron path from input to output within the target model.Our experiments demonstrate the superiority of our method finding the most influential neuron path along which the information flows, over the existing baseline solutions.Additionally, the neuron paths have illustrated that vision Transformers exhibit some specific inner working mechanism for processing the visual information within the same image category. We further analyze the key effects of these neurons on the image classification task, showcasing that the found neuron paths have already preserved the model capability on downstream tasks, which may also shed some lights on real-world applications like model pruning.The project website including implementation code is available at https://foundation-model-research.github.io/NeuronPath/.

1287. SGD with memory: fundamental properties and stochastic acceleration

链接：https://iclr.cc/virtual/2025/poster/29667 abstract： An important open problem is the theoretically feasible acceleration of mini-batch SGD-type algorithms on quadratic problems with power-law spectrum. In the non-stochastic setting, the optimal exponent ξ in the loss convergence $L_t\sim C_Lt^{-\xi}$ is double that in plain GD and is achievable using Heavy Ball (HB) with a suitable schedule; this no longer works in the presence of mini-batch noise. We address this challenge by considering first-order methods with an arbitrary fixed number M of auxiliary velocity vectors (*memory-M algorithms*). We first prove an equivalence between two forms of such algorithms and describe them in terms of suitable characteristic polynomials. Then we develop a general expansion of the loss in terms of *signal and noise propagators*. Using it, we show that losses of stationary stable memory-M algorithms always retain the exponent ξ of plain GD, but can have different constants C_L depending on their *effective learning rate* that generalizes that of HB. We prove that in memory-1 algorithms we can make C_L arbitrarily small while maintaining stability. As a consequence, we propose a memory-1 algorithm with a time-dependent schedule that we show heuristically and experimentally to improve the exponent ξ of plain SGD.

1288. Reconsidering Faithfulness in Regular, Self-Explainable and Domain Invariant GNNs

链接：https://iclr.cc/virtual/2025/poster/28564 abstract： As Graph Neural Networks (GNNs) become more pervasive, it becomes paramount to build reliable tools for explaining their predictions.A core desideratum is that explanations are faithful, i.e., that they portray an accurate picture of the GNN's reasoning process.However, a number of different faithfulness metrics exist, begging the question of what is faithfulness exactly and how to achieve it.We make three key contributions.We begin by showing that existing metrics are not interchangeable -- i.e., explanations attaining high faithfulness according to one metric may be unfaithful according to others -- and can systematically ignore important properties of explanations.We proceed to show that, surprisingly, optimizing for faithfulness is not always a sensible design goal. Specifically, we prove that for injective regular GNN architectures, perfectly faithful explanations are completely uninformative.This does not apply to modular GNNs, such as self-explainable and domain-invariant architectures, prompting us to study the relationship between architectural choices and faithfulness.Finally, we show that faithfulness is tightly linked to out-of-distribution generalization, in that simply ensuring that a GNN can correctly recognize the domain-invariant subgraph, as prescribed by the literature, does not guarantee that it is invariant unless this subgraph is also faithful.All our code can be found in the supplementary material.

1289. Beyond Surface Structure: A Causal Assessment of LLMs' Comprehension ability

链接：https://iclr.cc/virtual/2025/poster/28795 abstract： Large language models (LLMs) have shown remarkable capability in natural language tasks, yet debate persists on whether they truly comprehend deep structure (i.e., core semantics) or merely rely on surface structure (e.g., presentation format). Prior studies observe that LLMs' performance declines when intervening on surface structure, arguing their success relies on surface structure recognition. However, surface structure sensitivity does not prevent deep structure comprehension. Rigorously evaluating LLMs' capability requires analyzing both, yet deep structure is often overlooked. To this end, we assess LLMs' comprehension ability using causal mediation analysis, aiming to fully discover the capability of using both deep and surface structures. Specifically, we formulate the comprehension of deep structure as direct causal effect (DCE) and that of surface structure as indirect causal effect (ICE), respectively. To address the non-estimability of original DCE and ICE --- stemming from the infeasibility of isolating mutual influences of deep and surface structures, we develop the corresponding quantifiable surrogates, including approximated DCE (ADCE) and approximated ICE (AICE). We further apply the ADCE to evaluate a series of mainstream LLMs (and the one with random weights), showing that most of them exhibit deep structure comprehension ability, which grows along with the prediction accuracy. Comparing ADCE and AICE demonstrates closed-source LLMs (e.g., GPT) rely more on deep structure, while open-source LLMs (e.g., Llama) are more surface-sensitive, which decreases with model scale. Theoretically, ADCE is a bidirectional evaluation, which measures both the sufficiency and necessity of deep structure changes in causing output variations, thus offering a more comprehensive assessment than accuracy, a common evaluation in LLMs. Our work provides new insights into LLMs' deep structure comprehension and offers novel methods for LLMs evaluation. The code for our project is available at ADCE Project.

1290. OCCAM: Towards Cost-Efficient and Accuracy-Aware Classification Inference

链接：https://iclr.cc/virtual/2025/poster/30510 abstract： Classification tasks play a fundamental role in various applications, spanning domains such as healthcare, natural language processing and computer vision. With the growing popularity and capacity of machine learning models, people can easily access trained classifiers as a service online or offline. However, model use comes with a cost and classifiers of higher capacity (such as large foundation models) usually incur higher inference costs. To harness the respective strengths of different classifiers, we propose a principled approach, OCCAM, to compute the best classifier assignment strategy over classification queries (termed as the optimal model portfolio) so that the aggregated accuracy is maximized, under user-specified cost budgets. Our approach uses an unbiased and low-variance accuracy estimator and effectively computes the optimal solution by solving an integer linear programming problem. On a variety of real-world datasets, OCCAM achieves 40% cost reduction with little to no accuracy drop.

1291. StringLLM: Understanding the String Processing Capability of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28579 abstract： String processing, which mainly involves the analysis and manipulation of strings, is a fundamental component of modern computing. Despite the significant advancements of large language models (LLMs) in various natural language processing (NLP) tasks, their capability in string processing remains underexplored and underdeveloped. To bridge this gap, we present a comprehensive study of LLMs' string processing capability. In particular, we first propose StringLLM, a method to construct datasets for benchmarking string processing capability of LLMs. We use StringLLM to build a series of datasets, referred to as StringBench. It encompasses a wide range of string processing tasks, allowing us to systematically evaluate LLMs' performance in this area. Our evaluations indicate that LLMs struggle with accurately processing strings compared to humans. To uncover the underlying reasons for this limitation, we conduct an in-depth analysis and subsequently propose an effective approach that significantly enhances LLMs' string processing capability via fine-tuning. This work provides a foundation for future research to understand LLMs' string processing capability. Our code and data are available at https://github.com/wxl-lxw/StringLLM.

1292. Transformer-Squared: Self-adaptive LLMs

链接：https://iclr.cc/virtual/2025/poster/28974 abstract： Self-adaptive large language models (LLMs) aim to solve the challenges posed by traditional fine-tuning methods, which are often computationally intensive and static in their ability to handle diverse tasks. We introduce Transformer-Squared, a novel self-adaptation framework that adapts LLMs for unseen tasks in real-time by selectively adjusting only the singular components of their weight matrices. During inference, Transformer-Squared employs a two-pass mechanism: first, a dispatch system identifies the task properties, and then task-specific 'expert' vectors, trained using reinforcement learning, are dynamically mixed to obtain targeted behavior for the incoming prompt. Our method consistently outperforms ubiquitous approaches such as LoRA, with fewer parameters and greater efficiency. Furthermore, Transformer-Squared demonstrates versatility across different LLM architectures and modalities, including vision-language tasks. Transformer-Squared represents a significant leap forward, offering a scalable, efficient solution for enhancing the adaptability and task-specific performance of LLMs, paving the way for truly dynamic, self-organizing AI systems.

1293. Intrinsic Dimension Correlation: uncovering nonlinear connections in multimodal representations

链接：https://iclr.cc/virtual/2025/poster/29679 abstract： To gain insight into the mechanisms behind machine learning methods, it is crucial to establish connections among the features describing data points. However, these correlations often exhibit a high-dimensional and strongly nonlinear nature, which makes them challenging to detect using standard methods. This paper exploits the entanglement between intrinsic dimensionality and correlation to propose a metric that quantifies the (potentially nonlinear) correlation between high-dimensional manifolds. We first validate our method on synthetic data in controlled environments, showcasing its advantages and drawbacks compared to existing techniques. Subsequently, we extend our analysis to large-scale applications in neural network representations. Specifically, we focus on latent representations of multimodal data, uncovering clear correlations between paired visual and textual embeddings, whereas existing methods struggle significantly in detecting similarity. Our results indicate the presence of highly nonlinear correlation patterns between latent manifolds.

1294. Agent Skill Acquisition for Large Language Models via CycleQD

链接：https://iclr.cc/virtual/2025/poster/30026 abstract： Training large language models to acquire specific skills remains a challenging endeavor. Conventional training approaches often struggle with data distribution imbalances and inadequacies in objective functions that do not align well with task-specific performance. To address these challenges, we introduce CycleQD, a novel approach that leverages the Quality Diversity framework through a cyclic adaptation of the algorithm, along with a model merging based crossover and an SVD-based mutation. In CycleQD, each task’s performance metric is alternated as the quality measure while the others serve as the behavioral characteristics. This cyclic focus on individual tasks allows for concentrated effort on one task at a time, eliminating the need for data ratio tuning and simplifying the design of the objective function. Empirical results from AgentBench indicate that applying CycleQD to LLAMA3-8B-INSTRUCT based models not only enables them to surpass traditional fine-tuning methods in coding, operating systems, and database tasks, but also achieves performance on par with GPT-3.5-TURBO, which potentially contains much more parameters, across these domains. Crucially, this enhanced performance is achieved while retaining robust language capabilities, as evidenced by its performance on widely adopted language benchmark tasks. We highlight the key design choices in CycleQD, detailing how these contribute to its effectiveness. Furthermore, our method is general and can be applied to image segmentation models, highlighting its applicability across different domains.

1295. Controlling the Fidelity and Diversity of Deep Generative Models via Pseudo Density

链接：https://iclr.cc/virtual/2025/poster/31480 abstract： We introduce an approach to bias deep generative models, such as GANs and diffusion models, towards generating data with either enhanced fidelity or increased diversity. Our approach involves manipulating the distribution of training and generated data through a novel metric for individual samples, named pseudo density, which is based on the nearest-neighbor information from real samples. Our approach offers three distinct techniques to adjust the fidelity and diversity of deep generative models: 1) Per-sample perturbation, enabling precise adjustments for individual samples towards either more common or more unique characteristics; 2) Importance sampling during model inference to enhance either fidelity or diversity in the generated data; 3) Fine-tuning with importance sampling, which guides the generative model to learn an adjusted distribution, thus controlling fidelity and diversity. Furthermore, our fine-tuning method demonstrates the ability to improve the Frechet Inception Distance (FID) for pre-trained generative models with minimal iterations.

1296. Personalized Visual Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/32055 abstract： Recent advancements in multimodal large language models (MLLMs) have demonstrated significant progress; however, these models exhibit a notable limitation, which we refer to as "face blindness." Specifically, they can engage in general conversations but fail to conduct personalized dialogues targeting at specific individuals. This deficiency hinders the application of MLLMs in personalized settings, such as tailored visual assistants on mobile devices, or domestic robots that need to recognize members of the family. In this paper, we introduce Personalized Visual Instruction Tuning (PVIT), a novel data curation and training framework designed to enable MLLMs to identify target individuals within an image and engage in personalized and coherent dialogues. Our approach involves the development of a sophisticated pipeline that autonomously generates training data containing personalized conversations. This pipeline leverages the capabilities of various visual experts, image generation models, and (multi-modal) large language models. To evaluate the personalized potential of MLLMs, we present a benchmark called P-Bench, which encompasses various question types with different levels of difficulty. The experiments demonstrate a substantial personalized performance enhancement after fine-tuning with our curated dataset.

1297. PFGuard: A Generative Framework with Privacy and Fairness Safeguards

链接：https://iclr.cc/virtual/2025/poster/30740 abstract： Generative models must ensure both privacy and fairness for Trustworthy AI. While these goals have been pursued separately, recent studies propose to combine existing privacy and fairness techniques to achieve both goals. However, naively combining these techniques can be insufficient due to privacy-fairness conflicts, where a sample in a minority group may be represented in ways that support fairness, only to be suppressed for privacy. We demonstrate how these conflicts lead to adverse effects, such as privacy violations and unexpected fairness-utility tradeoffs. To mitigate these risks, we propose PFGuard, a generative framework with privacy and fairness safeguards, which simultaneously addresses privacy, fairness, and utility. By using an ensemble of multiple teacher models, PFGuard balances privacy-fairness conflicts between fair and private training stages and achieves high utility based on ensemble learning. Extensive experiments show that PFGuard successfully generates synthetic data on high-dimensional data while providing both DP guarantees and convergence in fair generative modeling.

1298. Hessian-Free Online Certified Unlearning

链接：https://iclr.cc/virtual/2025/poster/30533 abstract： Machine unlearning strives to uphold the data owners' right to be forgotten by enabling models to selectively forget specific data. Recent advances suggest pre-computing and storing statistics extracted from second-order information and implementing unlearning through Newton-style updates.However, the Hessian matrix operations are extremely costly and previous works conduct unlearning for empirical risk minimizer with the convexity assumption, precluding their applicability to high-dimensional over-parameterized models and the nonconvergence condition.In this paper, we propose an efficient Hessian-free unlearning approach. The key idea is to maintain a statistical vector for each training data, computed through affine stochastic recursion of the difference between the retrained and learned models. We prove that our proposed method outperforms the state-of-the-art methods in terms of the unlearning and generalization guarantees, the deletion capacity, and the time/storage complexity, under the same regularity conditions.Through the strategy of recollecting statistics for removing data, we develop an online unlearning algorithm that achieves near-instantaneous data removal, as it requires only vector addition.Experiments demonstrate that our proposed scheme surpasses existing results by orders of magnitude in terms of time/storage costs with millisecond-level unlearning execution, while also enhancing test accuracy.

1299. Integrative Decoding: Improving Factuality via Implicit Self-consistency

链接：https://iclr.cc/virtual/2025/poster/28829 abstract： Self-consistency-based approaches, which involve repeatedly sampling multiple outputs and selecting the most consistent one as the final response, prove to be remarkably effective in improving the factual accuracy of large language models. Nonetheless, existing methods usually have strict constraints on the task format, largely limiting their applicability. In this paper, we present Integrative Decoding (ID), to unlock the potential of self-consistency in open-ended generation tasks. ID operates by constructing a set of inputs, each prepended with a previously sampled response, and then processes them concurrently, with the next token being selected by aggregating of all their corresponding predictions at each decoding step. In essence, this simple approach implicitly incorporates self-consistency in the decoding objective. Extensive evaluation shows that ID consistently enhances factuality over a wide range of language models, with substantial improvements on the TruthfulQA (+11.2%), Biographies (+15.4%) and LongFact (+8.5%) benchmarks. The performance gains amplify progressively as the number of sampled responses increases, indicating the potential of ID to scale up with repeated sampling.

1300. Kolmogorov-Arnold Transformer

链接：https://iclr.cc/virtual/2025/poster/30584 abstract： Transformers stand as the cornerstone of mordern deep learning. Traditionally, these models rely on multi-layerperceptron (MLP) layers to mix the information between channels. In this paper, we introduce the Kolmogorov–ArnoldTransformer (KAT), a novel architecture that replaces MLP layers with Kolmogorov-Arnold Network (KAN) layers toenhance the expressiveness and performance of the model. Integrating KANs into transformers, however, is no easyfeat, especially when scaled up. Specifically, we identify three key challenges: (C1) Base function. The standard B-splinefunction used in KANs is not optimized for parallel computing on modern hardware, resulting in slower inference speeds.(C2) Parameter and Computation Inefficiency. KAN requires a unique function for each input-output pair, making thecomputation extremely large. (C3) Weight initialization. The initialization of weights in KANs is particularly challengingdue to their learnable activation functions, which are critical for achieving convergence in deep neural networks. Toovercome the aforementioned challenges, we propose three key solutions: (S1) Rational basis. We replace B-spline functionswith rational functions to improve compatibility with modern GPUs. By implementing this in CUDA, we achieve fastercomputations. (S2) Group KAN. We share the activation weights through a group of neurons, to reduce the computationalload without sacrificing performance. (S3) Variance-preserving initialization. We carefully initialize the activation weightsto make sure that the activation variance is maintained across layers. With these designs, KAT scales effectively and readilyoutperforms traditional MLP-based transformers. We demonstrate the advantages of KAT across various tasks, includingimage recognition, object detection, and semantic segmentation. It consistently enhances performance over the standardtransformer architectures of different model sizes.

1301. GraphBridge: Towards Arbitrary Transfer Learning in GNNs

链接：https://iclr.cc/virtual/2025/poster/28805 abstract： Graph neural networks (GNNs) are conventionally trained on a per-domain, per-task basis. It creates a significant barrier in transferring the acquired knowledge to different, heterogeneous data setups. This paper introduces GraphBridge, a novel framework to enable knowledge transfer across disparate tasks and domains in GNNs, circumventing the need for modifications to task configurations or graph structures. Specifically, GraphBridge allows for the augmentation of any pre-trained GNN with prediction heads and a bridging network that connects the input to the output layer. This architecture not only preserves the intrinsic knowledge of the original model but also supports outputs of arbitrary dimensions. To mitigate the negative transfer problem, GraphBridge merges the source model with a concurrently trained model, thereby reducing the source bias when applied to the target domain. Our method is thoroughly evaluated across diverse transfer learning scenarios, including Graph2Graph, Node2Node, Graph2Node, and graph2point-cloud. Empirical validation, conducted over 16 datasets representative of these scenarios, confirms the framework's capacity for task- and domain-agnostic transfer learning within graph-like data, marking a significant advancement in the field of GNNs. Code is available at https://github.com/jujulili888/GraphBridge.

1302. Factual Context Validation and Simplification: A Scalable Method to Enhance GPT Trustworthiness and Efficiency

链接：https://iclr.cc/virtual/2025/poster/31367 abstract： As the deployment of Large Language Models (LLMs) like GPT expands across domains, mitigating their susceptibility to factual inaccuracies or hallucinations becomes crucial for ensuring reliable performance. This blog post introduces two novel frameworks that enhance retrieval-augmented generation (RAG): one uses summarization to achieve a maximum of 57.7% storage reduction, while the other preserves critical information through statement-level extraction. Leveraging DBSCAN clustering, vectorized fact storage, and LLM-driven fact-checking, the pipelines deliver higher overall performance across benchmarks such as PubMedQA, SQuAD, and HotpotQA. By optimizing efficiency and accuracy, these frameworks advance trustworthy AI for impactful real-world applications.

1303. HG-Adapter: Improving Pre-Trained Heterogeneous Graph Neural Networks with Dual Adapters

链接：https://iclr.cc/virtual/2025/poster/30642 abstract： The "pre-train, prompt-tuning" paradigm has demonstrated impressive performance for tuning pre-trained heterogeneous graph neural networks (HGNNs) by mitigating the gap between pre-trained models and downstream tasks. However, most prompt-tuning-based works may face at least two limitations: (i) the model may be insufficient to fit the graph structures well as they are generally ignored in the prompt-tuning stage, increasing the training error to decrease the generalization ability; and (ii) the model may suffer from the limited labeled data during the prompt-tuning stage, leading to a large generalization gap between the training error and the test error to further affect the model generalization. To alleviate the above limitations, we first derive the generalization error bound for existing prompt-tuning-based methods, and then propose a unified framework that combines two new adapters with potential labeled data extension to improve the generalization of pre-trained HGNN models. Specifically, we design dual structure-aware adapters to adaptively fit task-related homogeneous and heterogeneous structural information. We further design a label-propagated contrastive loss and two self-supervised losses to optimize dual adapters and incorporate unlabeled nodes as potential labeled data. Theoretical analysis indicates that the proposed method achieves a lower generalization error bound than existing methods, thus obtaining superior generalization ability. Comprehensive experiments demonstrate the effectiveness and generalization of the proposed method on different downstream tasks.

1304. A General Framework for Producing Interpretable Semantic Text Embeddings

链接：https://iclr.cc/virtual/2025/poster/31167 abstract： Semantic text embedding is essential to many tasks in Natural Language Processing (NLP). While black-box models are capable of generating high-quality embeddings, their lack of interpretability limits their use in tasks that demand transparency. Recent approaches have improved interpretability by leveraging domain-expert-crafted or LLM-generated questions, but these methods rely heavily on expert input or well-prompt design, which restricts their generalizability and ability to generate discriminative questions across a wide range of tasks. To address these challenges, we introduce \algo{CQG-MBQA} (Contrastive Question Generation - Multi-task Binary Question Answering), a general framework for producing interpretable semantic text embeddings across diverse tasks. Our framework systematically generates highly discriminative, low cognitive load yes/no questions through the \algo{CQG} method and answers them efficiently with the \algo{MBQA} model, resulting in interpretable embeddings in a cost-effective manner. We validate the effectiveness and interpretability of \algo{CQG-MBQA} through extensive experiments and ablation studies, demonstrating that it delivers embedding quality comparable to many advanced black-box models while maintaining inherently interpretability. Additionally, \algo{CQG-MBQA} outperforms other interpretable text embedding methods across various downstream tasks. The source code is available at \url{https://github.com/dukesun99/CQG-MBQA}.

1305. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback

链接：https://iclr.cc/virtual/2025/poster/31506 abstract： Reinforcement learning from human feedback (RLHF) is a technique for training AI systems to align with human goals. RLHF has emerged as the central method used to finetune state-of-the-art large language models (LLMs). Despite this popularity, there has been relatively little public work systematizing its flaws. In this paper, we (1) survey open problems and fundamental limitations of RLHF and related methods; (2) overview techniques to understand, improve, and complement RLHF in practice; and (3) propose auditing and disclosure standards to improve societal oversight of RLHF systems. Our work emphasizes the limitations of RLHF and highlights the importance of a multi-layered approach to the development of safer AI systems.

1306. ColPali: Efficient Document Retrieval with Vision Language Models

链接：https://iclr.cc/virtual/2025/poster/28336 abstract： Documents are visually rich structures that convey information through text, but also figures, page layouts, tables, or even fonts. Since modern retrieval systems mainly rely on the textual information they extract from document pages to index documents -often through lengthy and brittle processes-, they struggle to exploit key visual cues efficiently. This limits their capabilities in many practical document retrieval applications such as Retrieval Augmented Generation (RAG).To benchmark current systems on visually rich document retrieval, we introduce the Visual Document Retrieval Benchmark $\textit{ViDoRe}$, composed of various page-level retrieval tasks spanning multiple domains, languages, and practical settings. The inherent complexity and performance shortcomings of modern systems motivate a new concept; doing document retrieval by directly embedding the images of the document pages. We release $\textit{ColPali}$, a Vision Language Model trained to produce high-quality multi-vector embeddings from images of document pages. Combined with a late interaction matching mechanism, $\textit{ColPali}$ largely outperforms modern document retrieval pipelines while being drastically simpler, faster and end-to-end trainable. We release models, data, code and benchmarks under open licenses at https://hf.co/vidore.

1307. NRGBoost: Energy-Based Generative Boosted Trees

链接：https://iclr.cc/virtual/2025/poster/27846 abstract： Despite the rise to dominance of deep learning in unstructured data domains, tree-based methods such as Random Forests (RF) and Gradient Boosted Decision Trees (GBDT) are still the workhorses for handling discriminative tasks on tabular data. We explore generative extensions of these popular algorithms with a focus on explicitly modeling the data density (up to a normalization constant), thus enabling other applications besides sampling. As our main contribution we propose an energy-based generative boosting algorithm that is analogous to the second-order boosting implemented in popular libraries like XGBoost. We show that, despite producing a generative model capable of handling inference tasks over any input variable, our proposed algorithm can achieve similar discriminative performance to GBDT on a number of real world tabular datasets, outperforming alternative generative approaches. At the same time, we show that it is also competitive with neural-network-based models for sampling.Code is available at https://github.com/ajoo/nrgboost.

1308. Neuron based Personality Trait Induction in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29990 abstract： Large language models (LLMs) have become increasingly proficient at simulating various personality traits, an important capability for supporting related applications (e.g., role-playing). To further improve this capacity, in this paper, we present a neuron based approach for personality trait induction in LLMs, with three major technical contributions. First, we construct PERSONALITYBENCH, a large-scale dataset for identifying and evaluating personality traits in LLMs. This dataset is grounded in the Big Five personality traits from psychology and designed to assess the generative capabilities of LLMs towards specific personality traits. Second, by leveraging PERSONALITYBENCH, we propose an efficient method for identifying personality-related neurons within LLMs by examining the opposite aspects of a given trait. Third, we develop a simple yet effective induction method that manipulates the values of these identified personality-related neurons, which enables fine-grained control over the traits exhibited by LLMs without training and modifying model parameters. Extensive experiments validates the efficacy of our neuron identification and trait induction methods. Notably, our approach achieves comparable performance as fine-tuned models, offering a more efficient and flexible solution for personality trait induction in LLMs.

1309. Support is All You Need for Certified VAE Training

链接：https://iclr.cc/virtual/2025/poster/28349 abstract： Variational Autoencoders (VAEs) have become increasingly popular and deployed in safety-critical applications. In such applications, we want to give certified probabilistic guarantees on performance under adversarial attacks. We propose a novel method, CIVET, for certified training of VAEs. CIVET depends on the key insight that we can bound worst-case VAE error by bounding the error on carefully chosen support sets at the latent layer. We show this point mathematically and present a novel training algorithm utilizing this insight. We show in an extensive evaluation across different datasets (in both the wireless and vision application areas), architectures, and perturbation magnitudes that our method outperforms SOTA methods achieving good standard performance with strong robustness guarantees.

1310. Robust Barycenter Estimation using Semi-Unbalanced Neural Optimal Transport

链接：https://iclr.cc/virtual/2025/poster/30524 abstract： Aggregating data from multiple sources can be formalized as an *Optimal Transport* (OT) barycenter problem, which seeks to compute the average of probability distributions with respect to OT discrepancies. However, in real-world scenarios, the presence of outliers and noise in the data measures can significantly hinder the performance of traditional statistical methods for estimating OT barycenters. To address this issue, we propose a novel scalable approach for estimating the *robust* continuous barycenter, leveraging the dual formulation of the *(semi-)unbalanced* OT problem. To the best of our knowledge, this paper is the first attempt to develop an algorithm for robust barycenters under the continuous distribution setup. Our method is framed as a $\min$-$\max$ optimization problem and is adaptable to *general* cost functions. We rigorously establish the theoretical underpinnings of the proposed method and demonstrate its robustness to outliers and class imbalance through a number of illustrative experiments. Our source code is publicly available at https://github.com/milenagazdieva/U-NOTBarycenters.

1311. SmartPretrain: Model-Agnostic and Dataset-Agnostic Representation Learning for Motion Prediction

链接：https://iclr.cc/virtual/2025/poster/30548 abstract： Predicting the future motion of surrounding agents is essential for autonomous vehicles (AVs) to operate safely in dynamic, human-robot-mixed environments. However, the scarcity of large-scale driving datasets has hindered the development of robust and generalizable motion prediction models, limiting their ability to capture complex interactions and road geometries. Inspired by recent advances in natural language processing (NLP) and computer vision (CV), self-supervised learning (SSL) has gained significant attention in the motion prediction community for learning rich and transferable scene representations. Nonetheless, existing pre-training methods for motion prediction have largely focused on specific model architectures and single dataset, limiting their scalability and generalizability.To address these challenges, we propose SmartPretrain, a general and scalable SSL framework for motion prediction that is both model-agnostic and dataset-agnostic. Our approach integrates contrastive and reconstructive SSL, leveraging the strengths of both generative and discriminative paradigms to effectively represent spatiotemporal evolution and interactions without imposing architectural constraints. Additionally, SmartPretrain employs a dataset-agnostic scenario sampling strategy that integrates multiple datasets, enhancing data volume, diversity, and robustness.Extensive experiments on multiple datasets demonstrate that SmartPretrain consistently improves the performance of state-of-the-art prediction models across datasets, data splits and main metrics. For instance, SmartPretrain significantly reduces the MissRate of Forecast-MAE by 10.6\%. These results highlight SmartPretrain's effectiveness as a unified, scalable solution for motion prediction, breaking free from the limitations of the small-data regime.

1312. Quamba: A Post-Training Quantization Recipe for Selective State Space Models

链接：https://iclr.cc/virtual/2025/poster/28449 abstract： State Space Models (SSMs) have emerged as an appealing alternative to Transformers for large language models, achieving state-of-the-art accuracy with constant memory complexity which allows for holding longer context lengths than attention-based networks. The superior computational efficiency of SSMs in long sequence modeling positions them favorably over Transformers in many scenarios. However, improving the efficiency of SSMs on request-intensive cloud-serving and resource-limited edge applications is still a formidable task. SSM quantization is a possible solution to this problem, making SSMs more suitable for wide deployment, while still maintaining their accuracy. Quantization is a common technique to reduce the model size and to utilize the low bit-width acceleration features on modern computing units, yet existing quantization techniques are poorly suited for SSMs. Most notably, SSMs have highly sensitive feature maps within the selective scan mechanism (i.e., linear recurrence) and massive outliers in the output activations which are not present in the output of token-mixing in the self-attention modules. To address this issue, we propose a static 8-bit per-tensor SSM quantization method which suppresses the maximum values of the input activations to the selective SSM for finer quantization precision and quantizes the output activations in an outlier-free space with Hadamard transform. Our 8-bit weight-activation quantized Mamba 2.8B SSM benefits from hardware acceleration and achieves a 1.72 $\times$ lower generation latency on an Nvidia Orin Nano 8G, with only a 0.9\% drop in average accuracy on zero-shot tasks. When quantizing Jamba, a 52B parameter SSM-style language model, we observe only a 1% drop in accuracy, demonstrating that our SSM quantization method is both effective and scalable for large language models, which require appropriate compression techniques for deployment. The experiments demonstrate the effectiveness and practical applicability of our approach for deploying SSM-based models of all sizes on both cloud and edge platforms.

1313. SWE-bench Multimodal: Do AI Systems Generalize to Visual Software

Domains?

链接：https://iclr.cc/virtual/2025/poster/28177 abstract： Autonomous systems for software engineering are now capable of fixing bugs and developing features. These systems are commonly evaluated on SWE-bench (Jimenez et al., 2024a), which assesses their ability to solve software issues from GitHub repositories. However, SWE-bench uses only Python repositories, with problem statements presented predominantly as text and lacking visual elements such as images. This limited coverage motivates our inquiry into how existing systems might perform on unrepresented software engineering domains(e.g., front-end, game development, DevOps), which use different programming languages and paradigms. Therefore, we propose SWE-bench Multimodal (SWE-bench M), to evaluate systems on their ability to fix bugs in visual, user-facing JavaScript software. SWE-bench M features 617 task instances collected from 17 JavaScript libraries used for web interface design, diagramming, data visualization, syntax highlighting, and interactive mapping. Each SWE-bench M task instance contains at least one image in its problem statement or unit tests. Our analysis finds that top-performing SWE-bench systems struggle with SWE-bench M, revealing limitations in visual problem-solving and cross-language generalization. Lastly, we show that SWE-agent's flexible language-agnostic features enable it to substantially outperform alternatives on SWE-bench M, resolving 12% of taskinstances compared to 6% for the next best system.

1314. Universal generalization guarantees for Wasserstein distributionally robust models

链接：https://iclr.cc/virtual/2025/poster/31241 abstract： Distributionally robust optimization has emerged as an attractive way to train robust machine learning models, capturing data uncertainty and distribution shifts. Recent statistical analyses have proved that generalization guarantees of robust models based on the Wasserstein distance have generalization guarantees that do not suffer from the curse of dimensionality. However, these results are either approximate, obtained in specific cases, or based on assumptions difficult to verify in practice. In contrast, we establish exact generalization guarantees that cover a wide range of cases, with arbitrary transport costs and parametric loss functions, including deep learning objectives with nonsmooth activations. We complete our analysis with an excess bound on the robust objective and an extension to Wasserstein robust models with entropic regularizations.

1315. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport

链接：https://iclr.cc/virtual/2025/poster/28002 abstract： Diffusion-based image generators are becoming unique methods for illumination harmonization and editing. The current bottleneck in scaling up the training of diffusion-based illumination editing models is mainly in the difficulty of preserving the underlying image details and maintaining intrinsic properties, such as albedos, unchanged. Without appropriate constraints, directly training the latest large image models with complex, varied, or in-the-wild data is likely to produce a structure-guided random image generator, rather than achieving the intended goal of precise illumination manipulation. We propose Imposing Consistent Light (IC-Light) transport during training, rooted in the physical principle that the linear blending of an object's appearances under different illumination conditions is consistent with its appearance under mixed illumination. This consistency allows for stable and scalable illumination learning, uniform handling of various data sources, and facilitates a physically grounded model behavior that modifies only the illumination of images while keeping other intrinsic properties unchanged. Based on this method, we can scale up the training of diffusion-based illumination editing models to large data quantities (> 10 million), across all available data types (real light stages, rendered samples, in-the-wild synthetic augmentations, etc), and using strong backbones (SDXL, Flux, etc). We also demonstrate that this approach reduces uncertainties and mitigates artifacts such as mismatched materials or altered albedos.

1316. Generating Graphs via Spectral Diffusion

链接：https://iclr.cc/virtual/2025/poster/30647 abstract： In this paper, we present GGSD, a novel graph generative model based on 1) the spectral decomposition of the graph Laplacian matrix and 2) a diffusion process. Specifically, we propose to use a denoising model to sample eigenvectors and eigenvalues from which we can reconstruct the graph Laplacian and adjacency matrix. Using the Laplacian spectrum allows us to naturally capture the structural characteristics of the graph and work directly in the node space while avoiding the quadratic complexity bottleneck that limits the applicability of other diffusion-based methods. This, in turn, is accomplished by truncating the spectrum, which, as we show in our experiments, results in a faster yet accurate generative process, and by designing a novel transformer-based architecture linear in the number of nodes. Our permutation invariant model can also handle node features by concatenating them to the eigenvectors of each node. An extensive set of experiments on both synthetic and real-world graphs demonstrates the strengths of our model against state-of-the-art alternatives.

1317. Sparse components distinguish visual pathways & their alignment to neural networks

链接：https://iclr.cc/virtual/2025/poster/30147 abstract： The ventral, dorsal, and lateral streams in high-level human visual cortex are implicated in distinct functional processes. Yet, deep neural networks (DNNs) trained on a single task model the entire visual system surprisingly well, hinting at common computational principles across these pathways. To explore this inconsistency, we applied a novel sparse decomposition approach to identify the dominant components of visual representations within each stream. Consistent with traditional neuroscience research, we find a clear difference in component response profiles across the three visual streams—identifying components selective for faces, places, bodies, text, and food in the ventral stream; social interactions, implied motion, and hand actions in the lateral stream; and some less interpretable components in the dorsal stream. Building on this, we introduce Sparse Component Alignment (SCA), a new method for measuring representational alignment between brains and machines that better captures the latent neural tuning of these two visual systems. We find that standard visual DNNs are more aligned with ventral than either dorsal or lateral representations. SCA reveals these distinctions with greater resolution than conventional population-level geometry, offering a measure of representational alignment that is sensitive to a system’s underlying axes of neural tuning.

1318. GaussianAnything: Interactive Point Cloud Flow Matching for 3D Generation

链接：https://iclr.cc/virtual/2025/poster/29781 abstract： Recent advancements in diffusion models and large-scale datasets have revolutionized image and video generation, with increasing focus on 3D content generation. While existing methods show promise, they face challenges in input formats, latent space structures, and output representations. This paper introduces a novel 3D generation framework that addresses these issues, enabling scalable and high-quality 3D generation with an interactive Point Cloud-structured Latent space. Our approach utilizes a VAE with multi-view posed RGB-D-N renderings as input, features a unique latent space design that preserves 3D shape information, and incorporates a cascaded latent flow-based model for improved shape-texture disentanglement. The proposed method, GaussianAnything, supports multi-modal conditional 3D generation, allowing for point cloud, caption, and single-view image inputs. Experimental results demonstrate superior performance on various datasets, advancing the state-of-the-art in 3D content generation.

1319. LongVILA: Scaling Long-Context Visual Language Models for Long Videos

链接：https://iclr.cc/virtual/2025/poster/27865 abstract： Long-context capability is critical for multi-modal foundation models, especially for long video understanding. We introduce LongVILA, a full-stack solution for long-context visual-language models by co-designing the algorithm and system. For model training, we upgrade existing VLMs to support long video understanding by incorporating two additional stages, i.e., long context extension and long video supervised fine-tuning. However, training on long video is computationally and memory intensive. We introduce the long-context Multi-Modal Sequence Parallelism (MM-SP) system that efficiently parallelizes long video training and inference, enabling 2M context length training on 256 GPUs without any gradient checkpointing. LongVILA efficiently extends the number of video frames of VILA from 8 to 2048, achieving 99.8% accuracy in 6,000-frame (more than 1 million tokens) video needle-in-a-haystack. LongVILA-7B demonstrates strong accuracy on 9 popular video benchmarks, e.g., 65.1% VideoMME with subtitle. Besides, MM-SP is 2.1x - 5.7x faster than ring style sequence parallelism and 1.1x - 1.4x faster than Megatron with a hybrid context and tensor parallelism. Moreover, it seamlessly integrates with Hugging Face Transformers.

1320. BirdSet: A Large-Scale Dataset for Audio Classification in Avian Bioacoustics

链接：https://iclr.cc/virtual/2025/poster/28986 abstract： Deep learning (DL) has greatly advanced audio classification, yet the field is limited by the scarcity of large-scale benchmark datasets that have propelled progress in other domains. While AudioSet is a pivotal step to bridge this gap as a universal-domain dataset, its restricted accessibility and limited range of evaluation use cases challenge its role as the sole resource. Therefore, we introduce BirdSet, a large-scale benchmark data set for audio classification focusing on avian bioacoustics. BirdSet surpasses AudioSet with over 6,800 recording hours ($\uparrow17\%$) from nearly 10,000 classes ($\uparrow18\times$) for training and more than 400 hours ($\uparrow7\times$) across eight strongly labeled evaluation datasets. It serves as a versatile resource for use cases such as multi-label classification, covariate shift or self-supervised learning. We benchmark six well-known DL models in multi-label classification across three distinct training scenarios and outline further evaluation use cases in audio classification. We host our dataset on Hugging Face for easy accessibility and offer an extensive codebase to reproduce our results.

1321. Eliciting Human Preferences with Language Models

链接：https://iclr.cc/virtual/2025/poster/29967 abstract： Language models (LMs) can be directed to perform user- and context-dependenttasks by using labeled examples or natural language prompts.But selecting examples or writing prompts can be challenging---especially in tasks that require users to precisely articulate nebulous preferences or reason about complex edge cases. For such tasks, we introduce Generative Active Task Elicitation (GATE), a method for using LMs themselves to guide the task specification process. GATE is a learning framework in which models elicit and infer human preferences through free-form, language-based interaction with users.We identify prototypical challenges that users face when specifying preferences, and design three preference modeling tasks to study these challenges:content recommendation, moral reasoning, and email validation.In preregistered experiments, we show that LMs that learn to perform these tasks using GATE (by interactively querying users with open-ended questions) obtain preference specifications that are more informative than user-written prompts or examples. GATE matches existing task specification methods in the moral reasoning task, and significantly outperforms them in the content recommendation and email validation tasks. Users additionally report that interactive task elicitation requires less effort than prompting or example labeling and surfaces considerations that they did not anticipate on their own. Our findings suggest that LM-driven elicitation can be a powerful tool for aligning models to complex human preferences and values.

1322. What Makes a Maze Look Like a Maze?

链接：https://iclr.cc/virtual/2025/poster/30138 abstract： A unique aspect of human visual understanding is the ability to flexibly interpret abstract concepts: acquiring lifted rules explaining what they symbolize, grounding them across familiar and unfamiliar contexts, and making predictions or reasoning about them. While off-the-shelf vision-language models excel at making literal interpretations of images (e.g., recognizing object categories such as tree branches), they still struggle to make sense of such visual abstractions (e.g., how an arrangement of tree branches may form the walls of a maze). To address this challenge, we introduce Deep Schema Grounding (DSG), a framework that leverages explicit structured representations of visual abstractions for grounding and reasoning. At the core of DSG are schemas—dependency graph descriptions of abstract concepts that decompose them into more primitive-level symbols. DSG uses large language models to extract schemas, then hierarchically grounds concrete to abstract components of the schema onto images with vision-language models. The grounded schema is used to augment visual abstraction understanding. We systematically evaluate DSG and different methods in reasoning on our new Visual Abstractions Benchmark, which consists of diverse, real-world images of abstract concepts and corresponding question-answer pairs labeled by humans. We show that DSG significantly improves the abstract visual reasoning performance of vision-language models, and is a step toward human-aligned understanding of visual abstractions.

1323. Identifying latent state transitions in non-linear dynamical systems

链接：https://iclr.cc/virtual/2025/poster/29017 abstract： This work aims to recover the underlying states and their time evolution in a latent dynamical system from high-dimensional sensory measurements. Previous works on identifiable representation learning in dynamical systems focused on identifying the latent states, often with linear transition approximations. As such, they cannot identify nonlinear transition dynamics, and hence fail to reliably predict complex future behavior. Inspired by the advances in nonlinear ICA, we propose a state-space modeling framework in which we can identify not just the latent states but also the unknown transition function that maps the past states to the present. Our identifiability theory relies on two key assumptions: (i) sufficient variability in the latent noise, and (ii) the bijectivity of the augmented transition function. Drawing from this theory, we introduce a practical algorithm based on variational auto-encoders. We empirically demonstrate that it improves generalization and interpretability of target dynamical systems by (i) recovering latent state dynamics with high accuracy, (ii) correspondingly achieving high future prediction accuracy, and (iii) adapting fast to new environments. Additionally, for complex real-world dynamics, (iv) it produces state-of the-art future prediction results for long horizons, highlighting its usefulness for practical scenarios.

1324. Selective Task Group Updates for Multi-Task Optimization

链接：https://iclr.cc/virtual/2025/poster/30394 abstract： Multi-task learning enables the acquisition of task-generic knowledge by training multiple tasks within a unified architecture. However, training all tasks together in a single architecture can lead to performance degradation, known as negative transfer, which is a main concern in multi-task learning. Previous works have addressed this issue by optimizing the multi-task network through gradient manipulation or weighted loss adjustments. However, their optimization strategy focuses on addressing task imbalance in shared parameters, neglecting the learning of task-specific parameters. As a result, they show limitations in mitigating negative transfer, since the learning of shared space and task-specific information influences each other during optimization. To address this, we propose a different approach to enhance multi-task performance by selectively grouping tasks and updating them for each batch during optimization. We introduce an algorithm that adaptively determines how to effectively group tasks and update them during the learning process. To track inter-task relations and optimize multi-task networks simultaneously, we propose proximal inter-task affinity, which can be measured during the optimization process. We provide a theoretical analysis on how dividing tasks into multiple groups and updating them sequentially significantly affects multi-task performance by enhancing the learning of task-specific parameters. Our methods substantially outperform previous multi-task optimization approaches and are scalable to different architectures and various numbers of tasks.

1325. TULIP: Token-length Upgraded CLIP

链接：https://iclr.cc/virtual/2025/poster/28217 abstract： We address the challenge of representing long captions in vision-language models, such as CLIP. By design these models are limited by fixed, absolute positional encodings, restricting inputs to a maximum of 77 tokens and hindering performance on tasks requiring longer descriptions. Although recent work has attempted to overcome this limit, their proposed approaches struggle to model token relationships over longer distances and simply extend to a fixed new token length. Instead, we propose a generalizable method, named TULIP, able to upgrade the token length to any length for CLIP-like models. We do so by improving the architecture with relative position encodings, followed by a training procedure that (i) distills the original CLIP text encoder into an encoder with relative position encodings and (ii) enhances the model for aligning longer captions with images. By effectively encoding captions longer than the default 77 tokens, our model outperforms baselines on cross-modal tasks such as retrieval and text-to-image generation. The code repository is available at https://github.com/ivonajdenkoska/tulip.

1326. From Few to Many: Self-Improving Many-Shot Reasoners Through

Iterative Optimization and Generation

链接：https://iclr.cc/virtual/2025/poster/30122 abstract： Recent advances in long-context large language models (LLMs) have led to the emerging paradigm of many-shot in-context learning (ICL), where it is observed that scaling many more demonstrating examples beyond the conventional few-shot setup in the context can lead to performance benefits. However, despite its promise, it is unclear what aspects dominate the benefits and whether simply scaling to more examples is the most effective way of improving many-shot ICL. In this work, we first provide an analysis on the factors driving many-shot ICL, and we find that 1) many-shot performance can still be attributed to often a few disproportionately influential examples and 2) identifying such influential examples ("optimize") and using them as demonstrations to regenerate new examples ("generate") can lead to further improvements. Inspired by the findings, we propose BRIDGE, an algorithm that alternates between the optimize step with Bayesian optimization to discover the influential sets of examples and the generate step to reuse this set to expand the reasoning paths of the examples back to the many-shot regime automatically. On Gemini, Claude, and Mistral LLMs of different sizes, we show BRIDGE led to significant improvements across a diverse set of tasks including symbolic reasoning, numerical reasoning and code generation.

1327. Real-time design of architectural structures with differentiable mechanics and neural networks

链接：https://iclr.cc/virtual/2025/poster/29515 abstract： Designing mechanically efficient geometry for architectural structures like shells, towers, and bridges, is an expensive iterative process.Existing techniques for solving such inverse problems rely on traditional optimization methods, which are slow and computationally expensive, limiting iteration speed and design exploration.Neural networks would seem to offer a solution via data-driven amortized optimization, but they often require extensive fine-tuning and cannot ensure that important design criteria, such as mechanical integrity, are met.In this work, we combine neural networks with a differentiable mechanics simulator to develop a model that accelerates the solution of shape approximation problems for architectural structures represented as bar systems.This model explicitly guarantees compliance with mechanical constraints while generating designs that closely match target geometries.We validate our approach in two tasks, the design of masonry shells and cable-net towers.Our model achieves better accuracy and generalization than fully neural alternatives, and comparable accuracy to direct optimization but in real time, enabling fast and reliable design exploration.We further demonstrate its advantages by integrating it into 3D modeling software and fabricating a physical prototype.Our work opens up new opportunities for accelerated mechanical design enhanced by neural networks for the built environment.

1328. Scalable Decentralized Learning with Teleportation

链接：https://iclr.cc/virtual/2025/poster/30604 abstract： Decentralized SGD can run with low communication costs, but its sparse communication characteristics deteriorate the convergence rate, especially when the number of nodes is large. In decentralized learning settings, communication is assumed to occur on only a given topology, while in many practical cases, the topology merely represents a preferred communication pattern, and connecting to arbitrary nodes is still possible. Previous studies have tried to alleviate the convergence rate degradation in these cases by designing topologies with large spectral gaps. However, the degradation is still significant when the number of nodes is substantial. In this work, we propose TELEPORTATION. TELEPORTATION activates only a subset of nodes, and the active nodes fetch the parameters from previous active nodes. Then, the active nodes update their parameters by SGD and perform gossip averaging on a relatively small topology comprising only the active nodes. We show that by activating only a proper number of nodes, TELEPORTATION can completely alleviate the convergence rate degradation. Furthermore, we propose an efficient hyperparameter-tuning method to search for the appropriate number of nodes to be activated. Experimentally, we showed that TELEPORTATION can train neural networks more stably and achieve higher accuracy than Decentralized SGD.

1329. PhiNets: Brain-inspired Non-contrastive Learning Based on Temporal Prediction Hypothesis

链接：https://iclr.cc/virtual/2025/poster/30926 abstract： Predictive coding has been established as a promising neuroscientific theory to describe the mechanism of information processing in the retina or cortex.This theory hypothesises that cortex predicts sensory inputs at various levels of abstraction to minimise prediction errors. Inspired by predictive coding, Chen et al. (2024) proposed another theory, temporal prediction hypothesis, to claim that sequence memory residing in hippocampus has emerged through predicting input signals from the past sensory inputs. Specifically, they supposed that the CA3 predictor in hippocampus creates synaptic delay between input signals, which is compensated by the following CA1 predictor. Though recorded neural activities were replicated based on the temporal prediction hypothesis, its validity has not been fully explored. In this work, we aim to explore the temporal prediction hypothesis from the perspective of self-supervised learning (SSL). Specifically, we focus on non-contrastive learning, which generates two augmented views of an input image and predicts one from another. Non-contrastive learning is intimately related to the temporal prediction hypothesis because the synaptic delay is implicitly created by StopGradient. Building upon a popular non-contrastive learner, SimSiam, we propose PhiNet, an extension of SimSiam to have two predictors explicitly corresponding to the CA3 and CA1, respectively. Through studying the PhiNet model, we discover two findings. First, meaningful data representations emerge in PhiNet more stably than in SimSiam. This is initially supported by our learning dynamics analysis: PhiNet is more robust to the representational collapse. Second, PhiNet adapts more quickly to newly incoming patterns in online and continual learning scenarios. For practitioners, we additionally propose an extension called X-PhiNet integrated with a momentum encoder, excelling in continual learning. All in all, our work reveals that the temporal prediction hypothesis is a reasonable model in terms of the robustness and adaptivity.

1330. Self-Supervised Diffusion Models for Electron-Aware Molecular Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29476 abstract： Physical properties derived from electronic distributions are essential information that determines molecular properties. However, the electron-level information is not accessible in most real-world complex molecules due to the extensive computational costs of determining uncertain electronic distributions. For this reason, existing methods for molecular property prediction have remained in regression models on simplified atom-level molecular descriptors, such as atomic structures and fingerprints. This paper proposes an efficient knowledge transfer method for electron-aware molecular representation learning. To this end, we devised a self-supervised diffusion method that estimates the electron-level information of real-world complex molecules without expensive quantum mechanical calculations. The proposed method achieved state-of-the-art prediction accuracy in the tasks of predicting molecular properties on extensive real-world molecular datasets.

1331. Learning Distributions of Complex Fluid Simulations with Diffusion Graph Networks

链接：https://iclr.cc/virtual/2025/poster/27985 abstract： Physical systems with complex unsteady dynamics, such as fluid flows, are often poorly represented by a single mean solution. For many practical applications, it is crucial to access the full distribution of possible states, from which relevant statistics (e.g., RMS and two-point correlations) can be derived. Here, we propose a graph-based latent diffusion model that enables direct sampling of states from their equilibrium distribution, given a mesh discretization of the system and its physical parameters. This allows for the efficient computation of flow statistics without running long and expensive numerical simulations. The graph-based structure enables operations on unstructured meshes, which is critical for representing complex geometries with spatially localized high gradients, while latent-space diffusion modeling with a multi-scale GNN allows for efficient learning and inference of entire distributions of solutions. A key finding of our work is that the proposed networks can accurately learn full distributions even when trained on incomplete data from relatively short simulations. We apply this method to a range of fluid dynamics tasks, such as predicting pressure distributions on 3D wing models in turbulent flow, demonstrating both accuracy and computational efficiency in challenging scenarios. The ability to directly sample accurate solutions, and capturing their diversity from short ground-truth simulations, is highly promising for complex scientific modeling tasks.

1332. Palu: KV-Cache Compression with Low-Rank Projection

链接：https://iclr.cc/virtual/2025/poster/29993 abstract： Post-training KV-Cache compression methods typically either sample a subset of effectual tokens or quantize the data into lower numerical bit width. However, these methods cannot exploit redundancy in the hidden dimension of the KV tenors. This paper presents a hidden dimension compression approach called Palu, a KV-Cache compression framework that utilizes low-rank projection to reduce inference-time LLM memory usage. Palu decomposes the linear layers into low-rank matrices, caches compressed intermediate states, and reconstructs the full keys and values on the fly. To improve accuracy, compression rate, and efficiency, Palu further encompasses (1) a medium-grained low-rank decomposition scheme, (2) an efficient rank search algorithm, (3) low-rank-aware quantization compatibility enhancements, and (4) an optimized GPU kernel with matrix fusion. Extensive experiments with popular LLMs show that Palu compresses KV-Cache by 50% while maintaining strong accuracy and delivering up to 1.89× speedup on the RoPE-based attention module. When combined with quantization, Palu’sinherent quantization-friendly design yields small to negligible extra accuracy degradation while saving additional memory than quantization-only methods and achieving up to 2.91× speedup for the RoPE-based attention. Moreover, it maintains comparable or even better accuracy (up to 1.19 lower perplexity) compared to quantization-only methods. These results demonstrate Palu’s superior capability to effectively address the efficiency and memory challenges of LLM inference posed by KV-Cache. Our code is publicly available at: https://github.com/shadowpa0327/Palu.

1333. REFINE: Inversion-Free Backdoor Defense via Model Reprogramming

链接：https://iclr.cc/virtual/2025/poster/31020 abstract： Backdoor attacks on deep neural networks (DNNs) have emerged as a significant security threat, allowing adversaries to implant hidden malicious behaviors during the model training phase. Pre-processing-based defense, which is one of the most important defense paradigms, typically focuses on input transformations or backdoor trigger inversion (BTI) to deactivate or eliminate embedded backdoor triggers during the inference process. However, these methods suffer from inherent limitations: transformation-based defenses often fail to balance model utility and defense performance, while BTI-based defenses struggle to accurately reconstruct trigger patterns without prior knowledge. In this paper, we propose REFINE, an inversion-free backdoor defense method based on model reprogramming. REFINE consists of two key components: \textbf{(1)} an input transformation module that disrupts both benign and backdoor patterns, generating new benign features; and \textbf{(2)} an output remapping module that redefines the model's output domain to guide the input transformations effectively. By further integrating supervised contrastive loss, REFINE enhances the defense capabilities while maintaining model utility. Extensive experiments on various benchmark datasets demonstrate the effectiveness of our REFINE and its resistance to potential adaptive attacks.

1334. To Trust or Not to Trust? Enhancing Large Language Models' Situated Faithfulness to External Contexts

链接：https://iclr.cc/virtual/2025/poster/30079 abstract： Large Language Models (LLMs) are often augmented with external contexts, such as those used in retrieval-augmented generation (RAG). However, these contexts can be inaccurate or intentionally misleading, leading to conflicts with the model’s internal knowledge. We argue that robust LLMs should demonstrate situated faithfulness, dynamically calibrating their trust in external information based on their confidence in the internal knowledge and the external context to resolve knowledge conflicts. To benchmark this capability, we evaluate LLMs across several QA datasets, including a newly created dataset featuring in-the-wild incorrect contexts sourced from Reddit posts. We show that when provided with both correct and incorrect contexts, both open-source and proprietary models tend to overly rely on external information, regardless of its factual accuracy. To enhance situated faithfulness, we propose two approaches: Self-Guided Confidence Reasoning (SCR) and Rule-Based Confidence Reasoning (RCR). SCR enables models to self-access the confidence of external information relative to their own internal knowledge to produce the most accurate answer. RCR, in contrast, extracts explicit confidence signals from the LLM and determines the final answer using predefined rules. Our results show that for LLMs with strong reasoning capabilities, such as GPT-4o and GPT-4o mini, SCR outperforms RCR, achieving improvements of up to 24.2\% over a direct input augmentation baseline. Conversely, for a smaller model like Llama-3-8B, RCR outperforms SCR. Fine-tuning SCR with our proposed Confidence Reasoning Direct Preference Optimization (CR-DPO) method improves performance on both seen and unseen datasets, yielding an average improvement of 8.9\% on Llama-3-8B. In addition to quantitative results, we offer insights into the relative strengths of SCR and RCR. Our findings highlight promising avenues for improving situated faithfulness in LLMs.

1335. Approaching Rate-Distortion Limits in Neural Compression with Lattice Transform Coding

链接：https://iclr.cc/virtual/2025/poster/29503 abstract： Neural compression has brought tremendous progress in designing lossy compressors with good rate-distortion (RD) performance at low complexity. Thus far, neural compression design involves transforming the source to a latent vector, which is then rounded to integers and entropy coded. While this approach has been shown to be optimal on a few specific sources, we show that it can be highly sub-optimal on synthetic sources whose intrinsic dimensionality is greater than one. With integer rounding in the latent space, the quantization regions induced by neural transformations, remain square-like and fail to match those of optimal vector quantization. We demonstrate that this phenomenon is due to the choice of scalar quantization in the latent space, and not the transform design. By employing lattice quantization instead, we propose Lattice Transform Coding (LTC) and show that it approximately recovers optimal vector quantization at reasonable complexity. On real-world sources, LTC improves upon standard neural compressors. LTC also provides a framework that can integrate structurally (near) optimal information-theoretic designs into lossy compression; examples include block coding, which yields coding gain over optimal one-shot coding and approaches the asymptotically-achievable rate-distortion function, as well as nested lattice quantization for low complexity fixed-rate coding.

1336. Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training

链接：https://iclr.cc/virtual/2025/poster/29616 abstract： Large language models (LLMs), optimized through human feedback, have rapidly emerged as a leading paradigm for developing intelligent conversational assistants. However, despite their strong performance across many benchmarks, LLM-based agents might still lack conversational skills such as disambiguation -- when they are faced with ambiguity, they often overhedge or implicitly guess users' true intents rather than asking clarification questions. Under task-specific settings, high-quality conversation samples are often limited, constituting a bottleneck for LLMs' ability to learn optimal dialogue action policies. We propose Action-Based Contrastive Self-Training (ACT), a quasi-online preference optimization algorithm based on Direct Preference Optimization (DPO), that enables data-efficient dialogue policy learning in multi-turn conversation modeling. We demonstrate ACT's efficacy under data-efficient tuning scenarios, even when there is no action label available, using multiple real-world conversational tasks: tabular-grounded question-answering, machine reading comprehension, and AmbigSQL, a novel task for disambiguating information-seeking requests for complex SQL generation towards data analysis agents. Additionally, we propose evaluating LLMs' ability to function as conversational agents by examining whether they can implicitly recognize and reason about ambiguity in conversation. ACT demonstrates substantial conversation modeling improvements over standard tuning approaches like supervised fine-tuning and DPO.

1337. Mitigating Object Hallucination in MLLMs via Data-augmented Phrase-level Alignment

链接：https://iclr.cc/virtual/2025/poster/27739 abstract： Despite their significant advancements, Multimodal Large Language Models(MLLMs) often generate factually inaccurate information, referred to as hallucination.In this work, we address object hallucinations in MLLMs, where informationis generated about an object not present in the input image. We introduce Data-augmentedPhrase-level Alignment (DPA), a novel loss which can be applied toinstruction-tuned off-the-shelf MLLMs to mitigate hallucinations, while preservingtheir general vision-language capabilities. To fine-tune MLLMs with DPA, we firstgenerate a set of 'hallucinated' and 'correct' response pairs through generative dataaugmentation by selectively altering the ground-truth information of the correctresponses at a phrase level. The DPA loss is then used to train MLLMs to reducethe likelihood of hallucinated phrases compared to the correct ones. Our thoroughevaluation on various benchmarks confirms the effectiveness of DPA in mitigatinghallucination while retaining the out-of-the-box performance of the MLLMs ongeneral tasks. For instance, MLLMs finetuned with DPA, which we refer to as HallucinationAttenuated Language and Vision Assistant (HALVA), improve F1 by upto 13.4% on hallucination visual question-answering and reduce the hallucinationrate by up to 4.2% on image description tasks.

1338. AI2TALE: An Innovative Information Theory-based Approach for Learning to Localize Phishing Attacks

链接：https://iclr.cc/virtual/2025/poster/31048 abstract： Phishing attacks remain a significant challenge for detection, explanation, and defense, despite over a decade of research on both technical and non-technical solutions. AI-based phishing detection methods are among the most effective approaches for defeating phishing attacks, providing predictions on the vulnerability label (i.e., phishing or benign) of data. However, they often lack intrinsic explainability, failing to identify the specific information that triggers the classification. To this end, we propose AI2TALE, an innovative deep learning-based approach for email (the most common phishing medium) phishing attack localization. Our method aims to not only predict the vulnerability label of the email data but also provide the capability to automatically learn and identify the most important and phishing-relevant information (i.e., sentences) in the phishing email data, offering useful and concise explanations for the identified vulnerability. Extensive experiments on seven diverse real-world email datasets demonstrate the capability and effectiveness of our method in selecting crucial information, enabling accurate detection and offering useful and concise explanations (via the most important and phishing-relevant information triggering the classification) for the vulnerability of phishing emails. Notably, our approach outperforms state-of-the-art baselines by 1.5% to 3.5% on average in Label-Accuracy and Cognitive-True-Positive metrics under a weakly supervised setting, where only vulnerability labels are used without requiring ground truth phishing information.

1339. E-Valuating Classifier Two-Sample Tests

链接：https://iclr.cc/virtual/2025/poster/31497 abstract： We introduce a powerful deep classifier two-sample test for high-dimensional data based on E-values, called E-C2ST. Our test combines ideas from existing work on split likelihood ratio tests and predictive independence tests. The resulting E-values are suitable for anytime-valid sequential two-sample tests. This feature allows for more effective use of data in constructing test statistics. Through simulations and real data applications, we empirically demonstrate that E-C2ST achieves enhanced statistical power by partitioning datasets into multiple batches, beyond the conventional two-split (training and testing) approach of standard two-sample classifier tests. This strategy increases the power of the test, while keeping the type I error well below the desired significance level.

1340. Bayesian Image Regression with Soft-thresholded Conditional Autoregressive Prior

链接：https://iclr.cc/virtual/2025/poster/28171 abstract： In the analysis of brain functional MRI (fMRI) data using regression models, Bayesian methods are highly valued for their flexibility and ability to quantify uncertainty. However, these methods face computational challenges in high-dimensional settings typical of brain imaging, and the often pre-specified correlation structures may not accurately capture the true spatial relationships within the brain. To address these issues, we develop a general prior specifically designed for regression models with large-scale imaging data. We introduce the Soft-Thresholded Conditional AutoRegressive (ST-CAR) prior, which reduces instability to pre-fixed correlation structures and provides inclusion probabilities to account for the uncertainty in choosing active voxels in the brain. We apply the ST-CAR prior to scalar-on-image (SonI) and image-on-scalar (IonS) regression models—both critical in brain imaging studies—and develop efficient computational algorithms using variational inference (VI) and stochastic subsampling techniques. Simulation studies demonstrate that the ST-CAR prior outperforms existing methods in identifying active brain regions with complex correlation patterns, while our VI algorithms offer superior computational performance. We further validate our approach by applying the ST-CAR to working memory fMRI data from the Adolescent Brain Cognitive Development (ABCD) study, highlighting its effectiveness in practical brain imaging applications.

1341. Explore Theory of Mind: program-guided adversarial data generation for theory of mind reasoning

链接：https://iclr.cc/virtual/2025/poster/31166 abstract： Do large language models (LLMs) have theory of mind? A plethora of papers and benchmarks have been introduced to evaluate if current models have been able to develop this key ability of social intelligence. However, all rely on limited datasets with simple patterns that can potentially lead to problematic blind spots in evaluation and an overestimation of model capabilities. We introduce ExploreToM, the first framework to allow large-scale generation of diverse and challenging theory of mind data for robust training and evaluation. Our approach leverages an A* search over a custom domain-specific language to produce complex story structures and novel, diverse, yet plausible scenarios to stress test the limits of LLMs. Our evaluation reveals that state-of-the-art LLMs, such as Llama-3.1-70B and GPT-4o, show accuracies as low as 0% and 9% on ExploreToM-generated data, highlighting the need for more robust theory of mind evaluation. As our generations are a conceptual superset of prior work, fine-tuning on our data yields a 27-point accuracy improvement on the classic ToMi benchmark (Le et al., 2019). ExploreToM also enables uncovering underlying skills and factors missing for models to show theory of mind, such as unreliable state tracking or data imbalances, which may contribute to models' poor performance on benchmarks.

1342. BEEM: Boosting Performance of Early Exit DNNs using Multi-Exit Classifiers as Experts

链接：https://iclr.cc/virtual/2025/poster/30371 abstract： Early Exit (EE) techniques have emerged as a means to reduce inference latency in Deep Neural Networks (DNNs). The latency improvement and accuracy in these techniques crucially depend on the criteria used to make exit decisions. We propose a new decision criterion BEEM where exit classifiers are treated as experts and aggregate their confidence scores. The confidence scores are aggregated only if neighbouring experts are consistent in prediction as the samples pass through them, thus capturing their ensemble effect. A sample exits when the aggregated confidence value exceeds a threshold. The threshold is set using the error rates of the intermediate exits aiming to surpass the performance of conventional DNN inference. Experimental results on the COCO dataset for Image captioning and GLUE datasets for various language tasks demonstrate that our method enhances the performance of state-of-the-art EE methods, achieving improvements in speed-up by a factor $1.5\times$ to $2.1\times$. When compared to the final layer, its accuracy is comparable in harder Image Captioning and improves in the easier language tasks. The source code is available at https://github.com/Div290/BEEM1/tree/main.

1343. Linear Recurrences Accessible to Everyone

链接：https://iclr.cc/virtual/2025/poster/31368 abstract： Investigating linear RNNs such as Mamba, can be challenging because they are currently not efficiently expressible in PyTorch. We propose the abstraction of linear recurrences to gain intuition for the computational structure of these emerging deep learning architectures. After deriving their parallel algorithm, we gradually build towards a simple template CUDA extension for PyTorch. We hope that making linear recurrences accessible to a wider audience inspires further research on linear-time sequence mixing.

1344. Varying Shades of Wrong: Aligning LLMs with Wrong Answers Only

链接：https://iclr.cc/virtual/2025/poster/28317 abstract： In the absence of abundant reliable annotations for challenging tasks and contexts, how can we expand the frontier of LLM capabilities with potentially wrong answers? We focus on two research questions: (1) Can LLMs generate reliable preferences among wrong options? And if so, (2) Would alignment with such wrong-over-wrong preferences be helpful? We employ methods based on self-consistency, token probabilities, and LLM-as-a-judge to elicit wrong-over-wrong preferences, and fine-tune language models with preference optimization approaches using these synthesized preferences. Extensive experiments with seven LLMs and eight datasets demonstrate that (1) LLMs do have preliminary capability in distinguishing various shades of wrong, achieving up to 20.9% higher performance than random guess; (2) Alignment with wrong-over-wrong preferences helps LLMs to produce less wrong and sometimes even outright correct answers, while improving overall model calibration. Code and data are publicly available at https://github.com/yaojh18/Varying-Shades-of-Wrong.

1345. The Ramanujan Library - Automated Discovery on the Hypergraph of Integer Relations

链接：https://iclr.cc/virtual/2025/poster/30374 abstract： Fundamental mathematical constants appear in nearly every field of science, from physics to biology. Formulas that connect different constants often bring great insight by hinting at connections between previously disparate fields.Discoveries of such relations, however, have remained scarce events, relying on sporadic strokes of creativity by human mathematicians.Recent developments of algorithms for automated conjecture generation have accelerated the discovery of formulas for specific constants.Yet, the discovery of connections between constants has not been addressed.In this paper, we present the first library dedicated to mathematical constants and their interrelations. This library can serve as a central repository of knowledge for scientists from different areas, and as a collaborative platform for development of new algorithms.The library is based on a new representation that we propose for organizing the formulas of mathematical constants: a hypergraph, with each node representing a constant and each edge representing a formula.Using this representation, we propose and demonstrate a systematic approach for automatically enriching this library using PSLQ, an integer relation algorithm based on QR decomposition and lattice construction. During its development and testing, our strategy led to the discovery of 75 previously unknown connections between constants, including a new formula for the `first continued fraction' constant C_1, novel formulas for natural logarithms, and new formulas connecting π and e.The latter formulas generalize a century-old relation between π and e by Ramanujan, which until now was considered a singular formula and is now found to be part of a broader mathematical structure. The code supporting this library is a public, open-source API that can serve researchers in experimental mathematics and other fields of science.

1346. PersonalLLM: Tailoring LLMs to Individual Preferences

链接：https://iclr.cc/virtual/2025/poster/32110 abstract： As LLMs become capable of complex tasks, there is growing potential for personalized interactions tailored to the subtle and idiosyncratic preferences of the user. We present a public benchmark, PersonalLLM, focusing on adapting LLMs to provide maximal benefits for a particular user. Departing from existing alignment benchmarks that implicitly assume uniform preferences, we curate open-ended prompts paired with many high-quality answers over which users would be expected to display heterogeneous latent preferences. Instead of persona prompting LLMs based on high-level attributes (e.g., user race or response length), which yields homogeneous preferences relative to humans, we develop a method that can simulate a large user base with diverse preferences from a set of pre-trained reward models. Our dataset and generated personalities offer an innovative testbed for developing personalization algorithms that grapple with continual data sparsity---few relevant feedback from the particular user---by leveraging historical data from other (similar) users. We explore basic in-context learning and meta-learning baselines to illustrate the utility of PersonalLLM and highlight the need for future methodological development.

1347. Language Imbalance Driven Rewarding for Multilingual Self-improving

链接：https://iclr.cc/virtual/2025/poster/30045 abstract： Large Language Models (LLMs) have achieved state-of-the-art performance across numerous tasks. However, these advancements have predominantly benefited "first-class" languages such as English and Chinese, leaving many other languages underrepresented. This imbalance, while limiting broader applications, generates a natural preference ranking between languages, offering an opportunity to bootstrap the multilingual capabilities of LLM in a self-improving manner. Thus, we propose $\textit{Language Imbalance Driven Rewarding}$, where the inherent imbalance between dominant and non-dominant languages within LLMs is leveraged as a reward signal. Iterative DPO training demonstrates that this approach not only enhances LLM performance in non-dominant languages but also improves the dominant language's capacity, thereby yielding an iterative reward signal. Fine-tuning Meta-Llama-3-8B-Instruct over two iterations of this approach results in continuous improvements in multilingual performance across instruction-following and arithmetic reasoning tasks, evidenced by an average improvement of 7.46\% win rate on the X-AlpacaEval leaderboard and 13.9\% accuracy on the MGSM benchmark. This work serves as an initial exploration, paving the way for multilingual self-improvement of LLMs.

1348. Balanced Neural ODEs: nonlinear model order reduction and Koopman operator approximations

链接：https://iclr.cc/virtual/2025/poster/28422 abstract： Variational Autoencoders (VAEs) are a powerful framework for learning latent representations of reduced dimensionality, while Neural ODEs excel in learning transient system dynamics. This work combines the strengths of both to generate fast surrogate models with adjustable complexity reacting on time-varying inputs signals. By leveraging the VAE’s dimensionality reduction using a non-hierarchical prior, our method adaptively assigns stochastic noise, naturally complementing known NeuralODE training enhancements and enabling probabilistictime series modeling. We show that standard Latent ODEs struggle with dimensionality reduction in systems with time-varying inputs. Our approach mitigates this by continuously propagating variational parameters through time, establishing fixed information channels in latent space. This results in a flexible and robust method that can learn different system complexities, e.g. deep neural networks orlinear matrices. Hereby, it enables efficient approximation of the Koopman operator without the need for predefining its dimensionality. As our method balances dimensionality reduction and reconstruction accuracy, we call it Balanced Neural ODE (B-NODE). We demonstrate the effectiveness of this methods on several academic and real-world test cases, e.g. a power plant or MuJoCo data.

1349. APE: Faster and Longer Context-Augmented Generation via Adaptive Parallel Encoding

链接：https://iclr.cc/virtual/2025/poster/27725 abstract： Context-augmented generation (CAG) techniques, including RAG and ICL, require the efficient combination of multiple contexts to generate responses to user queries. Directly inputting these contexts as a sequence introduces a considerable computational burden by re-encoding the combined selection of contexts for every request. To address this, we explore the promising potential of parallel encoding to independently pre-compute and cache each context's KV states. This approach enables the direct loading of cached states during inference while accommodating more contexts through position reuse across contexts. However, due to misalignments in attention distribution, directly applying parallel encoding results in a significant performance drop. To enable effective and efficient CAG, we propose Adaptive Parallel Encoding (**APE**), which brings shared prefix, attention temperature, and scaling factor to align the distribution of parallel encoding with sequential encoding. Results on RAG and ICL tasks demonstrate that APE can preserve 98\% and 93\% sequential encoding performance using the same inputs while outperforming parallel encoding by 3.6\% and 7.9\%, respectively. It also scales to many-shot CAG, effectively encoding hundreds of contexts in parallel. Efficiency evaluation shows that APE can achieve an end-to-end 4.5$\times$ speedup by reducing 28$\times$ prefilling time for a 128K-length context. The code is available at https://github.com/Infini-AI-Lab/APE.

1350. ManiSkill-HAB: A Benchmark for Low-Level Manipulation in Home Rearrangement Tasks

链接：https://iclr.cc/virtual/2025/poster/30874 abstract： High-quality benchmarks are the foundation for embodied AI research, enabling significant advancements in long-horizon navigation, manipulation and rearrangement tasks. However, as frontier tasks in robotics get more advanced, they require faster simulation speed, more intricate test environments, and larger demonstration datasets. To this end, we present MS-HAB, a holistic benchmark for low-level manipulation and in-home object rearrangement. First, we provide a GPU-accelerated implementation of the Home Assistant Benchmark (HAB). We support realistic low-level control and achieve over 3x the speed of prior magical grasp implementations at a fraction of the GPU memory usage. Second, we train extensive reinforcement learning (RL) and imitation learning (IL) baselines for future work to compare against. Finally, we develop a rule-based trajectory filtering system to sample specific demonstrations from our RL policies which match predefined criteria for robot behavior and safety. Combining demonstration filtering with our fast environments enables efficient, controlled data generation at scale.

1351. HiLo: A Learning Framework for Generalized Category Discovery Robust to Domain Shifts

链接：https://iclr.cc/virtual/2025/poster/31128 abstract： Generalized Category Discovery (GCD) is a challenging task in which, given a partially labelled dataset, models must categorize all unlabelled instances, regardless of whether they come from labelled categories or from new ones. In this paper, we challenge a remaining assumption in this task: that all images share the same domain. Specifically, we introduce a new task and method to handle GCD when the unlabelled data also contains images from different domains to the labelled set. Our proposed `HiLo' networks extract High-level semantic and Low-level domain features, before minimizing the mutual information between the representations. Our intuition is that the clusterings based on domain information and semantic information should be independent. We further extend our method with a specialized domain augmentation tailored for the GCD task, as well as a curriculum learning approach. Finally, we construct a benchmark from corrupted fine-grained datasets as well as a large-scale evaluation on DomainNet with real-world domain shifts, reimplementing a number of GCD baselines in this setting. We demonstrate that HiLo outperforms SoTA category discovery models by a large margin on all evaluations.

1352. More Experts Than Galaxies: Conditionally-Overlapping Experts with Biologically-Inspired Fixed Routing

链接：https://iclr.cc/virtual/2025/poster/31179 abstract： The evolution of biological neural systems has led to both modularity and sparse coding, which enables energy efficiency and robustness across the diversity of tasks in the lifespan. In contrast, standard neural networks rely on dense, non-specialized architectures, where all model parameters are simultaneously updated to learn multiple tasks, leading to interference. Current sparse neural network approaches aim to alleviate this issue but are hindered by limitations such as 1) trainable gating functions that cause representation collapse, 2) disjoint experts that result in redundant computation and slow learning, and 3) reliance on explicit input or task IDs that limit flexibility and scalability.In this paper we propose Conditionally Overlapping Mixture of ExperTs (COMET), a general deep learning method that addresses these challenges by inducing a modular, sparse architecture with an exponential number of overlapping experts. COMET replaces the trainable gating function used in Sparse Mixture of Experts with a fixed, biologically inspired random projection applied to individual input representations. This design causes the degree of expert overlap to depend on input similarity, so that similar inputs tend to share more parameters. This results in faster learning per update step and improved out-of-sample generalization. We demonstrate the effectiveness of COMET on a range of tasks, including image classification, language modeling, and regression, using several popular deep learning architectures.

1353. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model

链接：https://iclr.cc/virtual/2025/poster/29617 abstract： We introduce Transfusion, a recipe for training a multi-modal model over discrete and continuous data.Transfusion combines the language modeling loss function (next token prediction) with diffusion to train a single transformer over mixed-modality sequences.We pretrain multiple Transfusion models up to 7B parameters from scratch on a mixture of text and image data, establishing scaling laws with respect to a variety of uni- and cross-modal benchmarks.Our experiments show that Transfusion scales significantly better than quantizing images and training a language model over discrete image tokens.By introducing modality-specific encoding and decoding layers, we can further improve the performance of Transfusion models, and even compress each image to just 16 patches.We further demonstrate that scaling our Transfusion recipe to 7B parameters and 2T multi-modal tokens produces a model that can generate images and text on a par with similar scale diffusion models and language models, reaping the benefits of both worlds.

1354. Score Forgetting Distillation: A Swift, Data-Free Method for Machine Unlearning in Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28804 abstract： The machine learning community is increasingly recognizing the importance of fostering trust and safety in modern generative AI (GenAI) models. We posit machine unlearning (MU) as a crucial foundation for developing safe, secure, and trustworthy GenAI models. Traditional MU methods often rely on stringent assumptions and require access to real data. This paper introduces Score Forgetting Distillation (SFD), an innovative MU approach that promotes the forgetting of undesirable information in diffusion models by aligning the conditional scores of "unsafe" classes or concepts with those of "safe" ones. To eliminate the need for real data, our SFD framework incorporates a score-based MU loss into the score distillation objective of a pretrained diffusion model. This serves as a regularization term that preserves desired generation capabilities while enabling the production of synthetic data through a one-step generator. Our experiments on pretrained label-conditional and text-to-image diffusion models demonstrate that our method effectively accelerates the forgetting of target classes or concepts during generation, while preserving the quality of other classes or concepts. This unlearned and distilled diffusion not only pioneers a novel concept in MU but also accelerates the generation speed of diffusion models. Our experiments and studies on a range of diffusion models and datasets confirm that our approach is generalizable, effective, and advantageous for MU in diffusion models. Code is available at https://github.com/tqch/score-forgetting-distillation. (Warning: This paper contains sexually explicit imagery, discussions of pornography, racially-charged terminology, and other content that some readers may find disturbing, distressing, and/or offensive.)

1355. Provable weak-to-strong generalization via benign overfitting

链接：https://iclr.cc/virtual/2025/poster/30984 abstract： The classic teacher-student model in machine learning posits that a strong teacher supervises a weak student to improve the student's capabilities. We instead consider the inverted situation, where a weak teacher supervises a strong student with imperfect pseudolabels. This paradigm was recently brought forth by \citet{burns2023weak} and termed \emph{weak-to-strong generalization}. We theoretically investigate weak-to-strong generalization for binary and multilabel classification in a stylized overparameterized spiked covariance model with Gaussian covariates where the weak teacher's pseudolabels are asymptotically like random guessing. Under these assumptions, we provably identify two asymptotic phases of the strong student's generalization after weak supervision: (1) successful generalization and (2) random guessing. Our techniques should eventually extend to weak-to-strong multiclass classification. Towards doing so, we prove a tight lower tail inequality for the maximum of correlated Gaussians, which may be of independent interest. Understanding the multilabel setting reinforces the value of using logits for weak supervision when they are available.

1356. Non-Stationary Dueling Bandits Under a Weighted Borda Criterion

链接：https://iclr.cc/virtual/2025/poster/31451 abstract： In K-armed dueling bandits, the learner receives preference feedback between arms, and the regret of an arm is defined in terms of its suboptimality to a winner arm. The non-stationary variant of the problem, motivated by concerns of changing user preferences, has received recent interest (Saha and Gupta, 2022; Buening and Saha, 2023; Suk and Agarwal, 2023). The goal here is to design algorithms with low dynamic regret, ideally without foreknowledge of the amount of change. The notion of regret here is tied to a notion of winner arm, most typically taken to be a so-called Condorcet winner or a Borda winner. However, the aforementioned results mostly focus on the Condorcet winner. In comparison, the Borda version of this problem has received less attention which is the focus of this work. We establish the first optimal and adaptive dynamic regret upper bound $\tilde{O}(\tilde{L}^{1/3} K^{1/3} T^{2/3})$, where $\tilde{L}$ is the unknown number of significant Borda winner switches.

We also introduce a novel weighted Borda score framework which generalizes both the Borda and Condorcet problems. This framework surprisingly allows a Borda-style regret analysis of the Condorcet problem and establishes improved bounds over the theoretical state-of-art in regimes with a large number of arms or many spurious changes in Condorcet winner. Such a generalization was not known and could be of independent interest.

1357. Scalable Decision-Making in Stochastic Environments through Learned Temporal Abstraction

链接：https://iclr.cc/virtual/2025/poster/28299 abstract： Sequential decision-making in high-dimensional continuous action spaces, particularly in stochastic environments, faces significant computational challenges. We explore this challenge in the traditional offline RL setting, where an agent must learn how to make decisions based on data collected through a stochastic behavior policy. We present \textit{Latent Macro Action Planner} (L-MAP), which addresses this challenge by learning a set of temporally extended macro-actions through a state-conditional Vector Quantized Variational Autoencoder (VQ-VAE), effectively reducing action dimensionality. L-MAP employs a (separate) learned prior model that acts as a latent transition model and allows efficient sampling of plausible actions. During planning, our approach accounts for stochasticity in both the environment and the behavior policy by using Monte Carlo tree search (MCTS). In offline RL settings, including stochastic continuous control tasks, L-MAP efficiently searches over discrete latent actions to yield high expected returns.Empirical results demonstrate that L-MAP maintains low decision latency despite increased action dimensionality. Notably, across tasks ranging from continuous control with inherently stochastic dynamics to high-dimensional robotic hand manipulation, L-MAP significantly outperforms existing model-based methods and performs on par with strong model-free actor-critic baselines, highlighting the effectiveness of the proposed approach in planning in complex and stochastic environments with high-dimensional action spaces.

1358. Has the Deep Neural Network learned the Stochastic Process? An Evaluation Viewpoint

链接：https://iclr.cc/virtual/2025/poster/31133 abstract： This paper presents the first systematic study of evaluating Deep Neural Networks (DNNs) designed to forecast the evolution of stochastic complex systems. We show that traditional evaluation methods like threshold-based classification metrics and error-based scoring rules assess a DNN's ability to replicate the observed ground truth but fail to measure the DNN's learning of the underlying stochastic process. To address this gap, we propose a new evaluation criteria called Fidelity to Stochastic Process (F2SP), representing the DNN's ability to predict the system property Statistic-GT—the ground truth of the stochastic process—and introduce an evaluation metric that exclusively assesses F2SP. We formalize F2SP within a stochastic framework and establish criteria for validly measuring it. We formally show that Expected Calibration Error (ECE) satisfies the necessary condition for testing F2SP, unlike traditional evaluation methods. Empirical experiments on synthetic datasets, including wildfire, host-pathogen, and stock market models, demonstrate that ECE uniquely captures F2SP. We further extend our study to real-world wildfire data, highlighting the limitations of conventional evaluation and discuss the practical utility of incorporating F2SP into model assessment. This work offers a new perspective on evaluating DNNs modeling complex systems by emphasizing the importance of capturing underlying the stochastic process.

1359. Beyond Mere Token Analysis: A Hypergraph Metric Space Framework for Defending Against Socially Engineered LLM Attacks

链接：https://iclr.cc/virtual/2025/poster/28172 abstract： Recent jailbreak attempts on Large Language Models (LLMs) have shifted from algorithm-focused to human-like social engineering attacks, with persuasion-based techniques emerging as a particularly effective subset. These attacks evolve rapidly, demonstrate high creativity, and boast superior attack success rates. To combat such threats, we propose a promising approach to enhancing LLM safety by leveraging the underlying geometry of input prompt token embeddings using hypergraphs. This approach allows us to model the differences in information flow between benign and malicious LLM prompts.In our approach, each LLM prompt is represented as a metric hypergraph, forming a compact metric space. We then construct a higher-order metric space over these compact metric hypergraphs using the Gromov-Hausdorff distance as a generalized metric. Within this space of metric hypergraph spaces, our safety filter learns to classify between harmful and benign prompts. Our study presents theoretical guarantees on the classifier's generalization error for novel and unseen LLM input prompts. Extensive empirical evaluations demonstrate that our method significantly outperforms both existing state-of-the-art generic defense mechanisms and naive baselines. Notably, our approach also achieves comparable performance to specialized defenses against algorithm-focused attacks.

1360. Adversarial Search Engine Optimization for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28746 abstract： Large Language Models (LLMs) are increasingly used in applications where the model selects from competing third-party content, such as in LLM-powered search engines or chatbot plugins.In this paper, we introduce Preference Manipulation Attacks, a new class of attacks that manipulate an LLM's selections to favor the attacker. We demonstrate that carefully crafted website content or plugin documentations can trick an LLM to promote the attacker products and discredit competitors, thereby increasing user traffic and monetization (a form of adversarial Search Engine Optimization).We show this can lead to a prisoner's dilemma, where all parties are incentivized to launch attacks, but this collectively degrades the LLM's outputs for everyone. We demonstrate our attacks on production LLM search engines (Bing and Perplexity) and plugin APIs (for GPT-4 and Claude). As LLMs are increasingly used to rank third-party content, we expect Preference Manipulation Attacks to emerge as a significant threat.

1361. Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI

链接：https://iclr.cc/virtual/2025/poster/27708 abstract：

1362. OLMoE: Open Mixture-of-Experts Language Models

链接：https://iclr.cc/virtual/2025/poster/27777 abstract： We introduce OLMoE, a fully open, state-of-the-art language model leveraging sparse Mixture-of-Experts (MoE). OLMoE-1B-7B has 7 billion (B) parameters but uses only 1B per input token. We pretrain it on 5 trillion tokens and further adapt it to create OLMoE-1B-7B-Instruct. Our models outperform all available models with similar active parameters, even surpassing larger ones like Llama2-13B-Chat and DeepSeekMoE-16B. We present novel findings on MoE training, define and analyze new routing properties showing high specialization in our model, and open-source all our work: model weights, training data, code, and logs.

1363. Gramian Multimodal Representation Learning and Alignment

链接：https://iclr.cc/virtual/2025/poster/28847 abstract： Human perception integrates multiple modalities—such as vision, hearing, and language—into a unified understanding of the surrounding reality. While recent multimodal models have achieved significant progress by aligning pairs of modalities via contrastive learning, their solutions are unsuitable when scaling to multiple modalities. These models typically align each modality to a designated anchor without ensuring the alignment of all modalities with each other, leading to suboptimal performance in tasks requiring a joint understanding of multiple modalities. In this paper, we structurally rethink the pairwise conventional approach to multimodal learning and we present the novel Gramian Representation Alignment Measure (GRAM), which overcomes the above-mentioned limitations. GRAM learns and then aligns n modalities directly in the higher-dimensional space in which modality embeddings lie by minimizing the Gramian volume of the k-dimensional parallelotope spanned by the modality vectors, ensuring the geometric alignment of all modalities simultaneously. GRAM can replace cosine similarity in any downstream method, holding for 2 to n modalities and providing more meaningful alignment with respect to previous similarity measures. The novel GRAM-based contrastive loss function enhances the alignment of multimodal models in the higher-dimensional embedding space, leading to new state-of-the-art performance in downstream tasks such as video-audio-text retrieval and audio-video classification.

1364. Measuring Non-Adversarial Reproduction of Training Data in Large

Language Models

链接：https://iclr.cc/virtual/2025/poster/30966 abstract： Large language models memorize parts of their training data. Memorizing short snippets and facts is required to answer questions about the world and to be fluent in any language. But models have also been shown to reproduce long verbatim sequences of memorized text when prompted by a motivated adversary. In this work, we investigate an intermediate regime of memorization that we call non-adversarial reproduction, where we quantify the overlap between model responses and pretraining data when responding to natural and benign prompts. For a variety of innocuous prompt categories (e.g., writing a letter or a tutorial), we show that up to 15% of the text output by popular conversational language models overlaps with snippets from the Internet. In worst cases, we find generations where 100% of the content can be found exactly online. For the same tasks, we find that human-written text has far less overlap with Internet data. We further study whether prompting strategies can close this reproduction gap between models and humans. While appropriate prompting can reduce non-adversarial reproduction on average, we find that mitigating worst-case reproduction of training data requires stronger defenses—even for benign interactions.

1365. O(d/T) Convergence Theory for Diffusion Probabilistic Models under Minimal Assumptions

链接：https://iclr.cc/virtual/2025/poster/31027 abstract： Score-based diffusion models, which generate new data by learning to reverse a diffusion process that perturbs data from the target distribution into noise, have achieved remarkable success across various generative tasks. Despite their superior empirical performance, existing theoretical guarantees are often constrained by stringent assumptions or suboptimal convergence rates. In this paper, we establish a fast convergence theory for the denoising diffusion probabilistic model (DDPM), a widely used SDE-based sampler, under minimal assumptions. Our analysis shows that, provided ℓ_{2}-accurate estimates of the score functions, the total variation distance between the target and generated distributions is upper bounded by $O(d/T)$ (ignoring logarithmic factors), where d is the data dimensionality and T is the number of steps. This result holds for any target distribution with finite first-order moment. To our knowledge, this improves upon existing convergence theory for the DDPM sampler, while imposing minimal assumptions on the target data distribution and score estimates. This is achieved through a novel set of analytical tools that provides a fine-grained characterization of how the error propagates at each step of the reverse process.

1366. SINGAPO: Single Image Controlled Generation of Articulated Parts in Objects

链接：https://iclr.cc/virtual/2025/poster/29817 abstract： We address the challenge of creating 3D assets for household articulated objects from a single image.Prior work on articulated object creation either requires multi-view multi-state input, or only allows coarse control over the generation process.These limitations hinder the scalability and practicality for articulated object modeling.In this work, we propose a method to generate articulated objects from a single image.Observing the object in a resting state from an arbitrary view, our method generates an articulated object that is visually consistent with the input image.To capture the ambiguity in part shape and motion posed by a single view of the object, we design a diffusion model that learns the plausible variations of objects in terms of geometry and kinematics.To tackle the complexity of generating structured data with attributes in multiple domains, we design a pipeline that produces articulated objects from high-level structure to geometric details in a coarse-to-fine manner, where we use a part connectivity graph and part abstraction as proxies.Our experiments show that our method outperforms the state-of-the-art in articulated object creation by a large margin in terms of the generated object realism, resemblance to the input image, and reconstruction quality.

1367. Hierarchical Autoregressive Transformers: Combining Byte- and Word-Level Processing for Robust, Adaptable Language Models

链接：https://iclr.cc/virtual/2025/poster/28049 abstract： Tokenization is a fundamental step in natural language processing, breaking text into units that computational models can process. While learned subword tokenizers have become the de-facto standard, they present challenges such as large vocabularies, limited adaptability to new domains or languages, and sensitivity to spelling errors and variations. To overcome these limitations, we investigate a hierarchical architecture for autoregressive language modelling that combines character-level and word-level processing. It employs a lightweight character-level encoder to convert character sequences into word embeddings, which are then processed by a word-level backbone model and decoded back into characters via a compact character-level decoder. This method retains the sequence compression benefits of word-level tokenization without relying on a rigid, predefined vocabulary. We demonstrate, at scales up to 7 billion parameters, that hierarchical transformers match the downstream task performance of subword-tokenizer-based models while exhibiting significantly greater robustness to input perturbations. Additionally, during continued pretraining on an out-of-domain language, our model trains almost twice as fast, achieves superior performance on the target language, and retains more of its previously learned knowledge. Hierarchical transformers pave the way for NLP systems that are more robust, flexible, and generalizable across languages and domains.

1368. One Model Transfer to All: On Robust Jailbreak Prompts Generation against LLMs

链接：https://iclr.cc/virtual/2025/poster/28127 abstract： Safety alignment in large language models (LLMs) is increasingly compromised by jailbreak attacks, which can manipulate these models to generate harmful or unintended content. Investigating these attacks is crucial for uncovering model vulnerabilities. However, many existing jailbreak strategies fail to keep pace with the rapid development of defense mechanisms, such as defensive suffixes, rendering them ineffective against defended models. To tackle this issue, we introduce a novel attack method called ArrAttack, specifically designed to target defended LLMs. ArrAttack automatically generates robust jailbreak prompts capable of bypassing various defense measures. This capability is supported by a universal robustness judgment model that, once trained, can perform robustness evaluation for any target model with a wide variety of defenses. By leveraging this model, we can rapidly develop a robust jailbreak prompt generator that efficiently converts malicious input prompts into effective attacks. Extensive evaluations reveal that ArrAttack significantly outperforms existing attack strategies, demonstrating strong transferability across both white-box and black-box models, including GPT-4 and Claude-3. Our work bridges the gap between jailbreak attacks and defenses, providing a fresh perspective on generating robust jailbreak prompts.

1369. IRIS: LLM-Assisted Static Analysis for Detecting Security Vulnerabilities

链接：https://iclr.cc/virtual/2025/poster/30698 abstract： Software is prone to security vulnerabilities. Program analysis tools to detect them have limited effectiveness in practice due to their reliance on human labeled specifications. Large language models (or LLMs) have shown impressive code generation capabilities but they cannot do complex reasoning over code to detect such vulnerabilities especially since this task requires whole-repository analysis. We propose IRIS, a neuro-symbolic approach that systematically combines LLMs with static analysis to perform whole-repository reasoning for security vulnerability detection. Specifically, IRIS leverages LLMs to infer taint specifications and perform contextual analysis, alleviating needs for human specifications and inspection. For evaluation, we curate a new dataset, CWE-Bench-Java, comprising 120 manually validated security vulnerabilities in real-world Java projects. A state-of-the-art static analysis tool CodeQL detects only 27 of these vulnerabilities whereas IRIS with GPT-4 detects 55 (+28) and improves upon CodeQL's average false discovery rate by 5% points.Furthermore, IRIS identifies 4 previously unknown vulnerabilities which cannot be found by existing tools. IRIS is available publicly at https://github.com/iris-sast/iris.

1370. Scalable Extraction of Training Data from Aligned, Production Language Models

链接：https://iclr.cc/virtual/2025/poster/27894 abstract： Large language models are prone to memorizing some of their training data. Memorized (and possibly sensitive) samples can then be extracted at generation time by adversarial or benign users. There is hope that model alignment---a standard training process that tunes a model to harmlessly follow user instructions---would mitigate the risk of extraction. However, we develop two novel attacks that undo a language model's alignment and recover thousands of training examples from popular proprietary aligned models such as OpenAI's ChatGPT. Our work highlights the limitations of existing safeguards to prevent training data leakage in production language models.

1371. Reti-Diff: Illumination Degradation Image Restoration with Retinex-based Latent Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/28550 abstract： Illumination degradation image restoration (IDIR) techniques aim to improve the visibility of degraded images and mitigate the adverse effects of deteriorated illumination. Among these algorithms, diffusion-based models (DM) have shown promising performance but are often burdened by heavy computational demands and pixel misalignment issues when predicting the image-level distribution. To tackle these problems, we propose to leverage DM within a compact latent space to generate concise guidance priors and introduce a novel solution called Reti-Diff for the IDIR task. Specifically, Reti-Diff comprises two significant components: the Retinex-based latent DM (RLDM) and the Retinex-guided transformer (RGformer). RLDM is designed to acquire Retinex knowledge, extracting reflectance and illumination priors to facilitate detailed reconstruction and illumination correction. RGformer subsequently utilizes these compact priors to guide the decomposition of image features into their respective reflectance and illumination components. Following this, RGformer further enhances and consolidates these decomposed features, resulting in the production of refined images with consistent content and robustness to handle complex degradation scenarios. Extensive experiments demonstrate that Reti-Diff outperforms existing methods on three IDIR tasks, as well as downstream applications.

1372. Overcoming False Illusions in Real-World Face Restoration with Multi-Modal Guided Diffusion Model

链接：https://iclr.cc/virtual/2025/poster/28479 abstract： We introduce a novel Multi-modal Guided Real-World Face Restoration (MGFR) technique designed to improve the quality of facial image restoration from low-quality inputs. Leveraging a blend of attribute text prompts, high-quality reference images, and identity information, MGFR can mitigate the generation of false facial attributes and identities often associated with generative face restoration methods. By incorporating a dual-control adapter and a two-stage training strategy, our method effectively utilizes multi-modal prior information for targeted restoration tasks. We also present the Reface-HQ dataset, comprising over 21,000 high-resolution facial images across 4800 identities, to address the need for reference face training images. Our approach achieves superior visual quality in restoring facial details under severe degradation and allows for controlled restoration processes, enhancing the accuracy of identity preservation and attribute correction. Including negative quality samples and attribute prompts in the training further refines the model's ability to generate detailed and perceptually accurate images.

1373. Gap Preserving Distillation by Building Bidirectional Mappings with A Dynamic Teacher

链接：https://iclr.cc/virtual/2025/poster/29731 abstract： Knowledge distillation aims to transfer knowledge from a large teacher model to a compact student counterpart, often coming with a significant performance gap between them. Interestingly, we find that a too-large performance gap can hamper the training process.To alleviate this, we propose a Gap Preserving Distillation (GPD) method that trains an additional dynamic teacher model from scratch along with the student to maintain a reasonable performance gap. To further strengthen distillation, we develop a hard strategy by enforcing both models to share parameters. Besides, we also build the soft bidirectional mappings between them through Inverse Reparameterization (IR) and Channel-Branch Reparameterization (CBR).IR initializes a larger dynamic teacher with approximately the same accuracy as the student to avoid a too large gap in early stage of training. CBR enables direct extraction of an effective student model from the dynamic teacher without post-training. In experiments, GPD significantly outperforms existing distillation methods on top of both CNNs and transformers, achieving up to 1.58\% accuracy improvement. Interestingly, GPD also generalizes well to the scenarios without a pre-trained teacher, including training from scratch and fine-tuning, yielding a large improvement of 1.80\% and 0.89\% on ResNet18, respectively.

1374. GenDataAgent: On-the-fly Dataset Augmentation with Synthetic Data

链接：https://iclr.cc/virtual/2025/poster/29344 abstract： We propose a generative agent that augments training datasets with synthetic data for model fine-tuning. Unlike prior work, which uniformly samples synthetic data, our agent iteratively generates relevant samples on-the-fly, aligning with the target distribution. It prioritizes synthetic data that complements difficult training samples, focusing on those with high variance in gradient updates. Experiments across several image classification tasks demonstrate the effectiveness of our approach.

1375. Sylber: Syllabic Embedding Representation of Speech from Raw Audio

链接：https://iclr.cc/virtual/2025/poster/30312 abstract： Syllables are compositional units of spoken language that efficiently structure human speech perception and production. However, current neural speech representations lack such structure, resulting in dense token sequences that are costly to process. To bridge this gap, we propose a new model, Sylber, that produces speech representations with clean and robust syllabic structure. Specifically, we propose a self-supervised learning (SSL) framework that bootstraps syllabic embeddings by distilling from its own initial unsupervised syllabic segmentation. This results in a highly structured representation of speech features, offering three key benefits: 1) a fast, linear-time syllable segmentation algorithm, 2) efficient syllabic tokenization with an average of 4.27 tokens per second, and 3) novel phonological units suited for efficient spoken language modeling. Our proposed segmentation method is highly robust and generalizes to out-of-domain data and unseen languages without any tuning. By training token-to-speech generative models, fully intelligible speech can be reconstructed from Sylber tokens with a significantly lower bitrate than baseline SSL tokens. This suggests that our model effectively compresses speech into a compact sequence of tokens with minimal information loss. Lastly, we demonstrate that categorical perception—a linguistic phenomenon in speech perception—emerges naturally in Sylber, making the embedding space more categorical and sparse than previous speech features and thus supporting the high efficiency of our tokenization. Together, we present a novel SSL approach for representing speech as syllables, with significant potential for efficient speech tokenization and spoken language modeling.

1376. Learning Interpretable Hierarchical Dynamical Systems Models from Time Series Data

链接：https://iclr.cc/virtual/2025/poster/29398 abstract： In science, we are often interested in obtaining a generative model of the underlying system dynamics from observed time series. While powerful methods for dynamical systems reconstruction (DSR) exist when data come from a single domain, how to best integrate data from multiple dynamical regimes and leverage it for generalization is still an open question. This becomes particularly important when individual time series are short, and group-level information may help to fill in for gaps in single-domain data. Here we introduce a hierarchical framework that enables to harvest group-level (multi-domain) information while retaining all single-domain characteristics, and showcase it on popular DSR benchmarks, as well as on neuroscience and medical data. In addition to faithful reconstruction of all individual dynamical regimes, our unsupervised methodology discovers common low-dimensional feature spaces in which datasets with similar dynamics cluster. The features spanning these spaces were further dynamically highly interpretable, surprisingly in often linear relation to control parameters that govern the dynamics of the underlying system. Finally, we illustrate transfer learning and generalization to new parameter regimes, paving the way toward DSR foundation models.

1377. CAT-3DGS: A Context-Adaptive Triplane Approach to Rate-Distortion-Optimized 3DGS Compression

链接：https://iclr.cc/virtual/2025/poster/28486 abstract： 3D Gaussian Splatting (3DGS) has recently emerged as a promising 3D representation. Much research has been focused on reducing its storage requirements and memory footprint. However, the needs to compress and transmit the 3DGS representation to the remote side are overlooked. This new application calls for rate-distortion-optimized 3DGS compression. How to quantize and entropy encode sparse Gaussian primitives in the 3D space remains largely unexplored. Few early attempts resort to the hyperprior framework from learned image compression. But, they fail to utilize fully the inter and intra correlation inherent in Gaussian primitives. Built on ScaffoldGS, this work, termed CAT-3DGS, introduces a context-adaptive triplane approach to their rate-distortion-optimized coding. It features multi-scale triplanes, oriented according to the principal axes of Gaussian primitives in the 3D space, to capture their inter correlation (i.e. spatial correlation) for spatial autoregressive coding in the projected 2D planes. With these triplanes serving as the hyperprior, we further perform channel-wise autoregressive coding to leverage the intra correlation within each individual Gaussian primitive. Our CAT-3DGS incorporates a view frequency-aware masking mechanism. It actively skips from coding those Gaussian primitives that potentially have little impact on the rendering quality. When trained end-to-end to strike a good rate-distortion trade-off, our CAT-3DGS achieves the state-of-the-art compression performance on the commonly used real-world datasets.

1378. Causal Graphical Models for Vision-Language Compositional Understanding

链接：https://iclr.cc/virtual/2025/poster/28753 abstract： Recent work has empirically shown that Vision-Language Models (VLMs) struggleto fully understand the compositional properties of the human language, usuallymodeling an image caption as a “bag of words”. As a result, they performpoorly on compositional tasks, which require a deeper understanding of the differententities of a sentence (subject, verb, etc.) jointly with their mutual relationshipsin order to be solved. In this paper, we model the dependency relationsamong textual and visual tokens using a Causal Graphical Model (CGM), built usinga dependency parser, and we train a decoder conditioned by the VLM visualencoder. Differently from standard autoregressive or parallel predictions, our decoder’sgenerative process is partially-ordered following the CGM structure. Thisstructure encourages the decoder to learn only the main causal dependencies ina sentence discarding spurious correlations. Using extensive experiments on fivecompositional benchmarks, we show that our method significantly outperformsall the state-of-the-art compositional approaches by a large margin, and it also improvesover methods trained using much larger datasets. Our model weights and code are publicly available.

1379. A Unified Theory of Quantum Neural Network Loss Landscapes

链接：https://iclr.cc/virtual/2025/poster/28843 abstract： Classical neural networks with random initialization famously behave as Gaussian processes in the limit of many neurons, which allows one to completely characterize their training and generalization behavior. No such general understanding exists for quantum neural networks (QNNs), which—outside of certain special cases—are known to not behave as Gaussian processes when randomly initialized. We here prove that QNNs and their first two derivatives instead generally form what we call "Wishart processes," where certain algebraic properties of the network determine the hyperparameters of the process. This Wishart process description allows us to, for the first time: give necessary and sufficient conditions for a QNN architecture to have a Gaussian process limit; calculate the full gradient distribution, generalizing previously known barren plateau results; and calculate the local minima distribution of algebraically constrained QNNs. Our unified framework suggests a certain simple operational definition for the "trainability" of a given QNN model using a newly introduced, experimentally accessible quantity we call the "degrees of freedom" of the network architecture.

1380. Emergence of meta-stable clustering in mean-field transformer models

链接：https://iclr.cc/virtual/2025/poster/28944 abstract： We model the evolution of tokens within a deep stack of Transformer layers as a continuous-time flow on the unit sphere, governed by a mean-field interacting particle system, building on the framework introduced in Geshkovski et al. (2023). Studying the corresponding mean-field Partial Differential Equation (PDE), which can be interpreted as a Wasserstein gradient flow, in this paper we provide a mathematical investigation of the long-term behavior of this system, with a particular focus on the emergence and persistence of meta-stable phases and clustering phenomena, key elements in applications like next-token prediction. More specifically, we perform a perturbative analysis of the mean-field PDE around the iid uniform initialization and prove that, in the limit of large number of tokens, the model remains close to a meta-stable manifold of solutions with a given structure (e.g., periodicity). Further, the structure characterizing the meta-stable manifold is explicitly identified, as a function of the inverse temperature parameter of the model, by the index maximizing a certain rescaling of Gegenbauer polynomials.

1381. Counterfactual Generative Modeling with Variational Causal Inference

链接：https://iclr.cc/virtual/2025/poster/28343 abstract： Estimating an individual's potential outcomes under counterfactual treatments is a challenging task for traditional causal inference and supervised learning approaches when the outcome is high-dimensional (e.g. gene expressions, facial images) and covariates are relatively limited. In this case, to predict one's outcomes under counterfactual treatments, it is crucial to leverage individual information contained in the observed outcome in addition to the covariates. Prior works using variational inference in counterfactual generative modeling have been focusing on neural adaptations and model variants within the conditional variational autoencoder formulation, which we argue is fundamentally ill-suited to the notion of counterfactual in causal inference. In this work, we present a novel variational Bayesian causal inference framework and its theoretical backings to properly handle counterfactual generative modeling tasks, through which we are able to conduct counterfactual supervision end-to-end during training without any counterfactual samples, and encourage disentangled exogenous noise abduction that aids the correct identification of causal effect in counterfactual generations. In experiments, we demonstrate the advantage of our framework compared to state-of-the-art models in counterfactual generative modeling on multiple benchmarks.

1382. Bias Mitigation in Graph Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/30512 abstract： Most existing graph diffusion models have significant bias problems. We observe that the forward diffusion’s maximum perturbation distribution in most models deviates from the standard Gaussian distribution, while reverse sampling consistently starts from a standard Gaussian distribution, which results in a reverse-starting bias. Together with the inherent exposure bias of diffusion models, this results in degraded generation quality. This paper proposes a comprehensive approach to mitigate both biases. To mitigate reverse-starting bias, we employ a newly designed Langevin sampling algorithm to align with the forward maximum perturbation distribution, establishing a new reverse-starting point. To address the exposure bias, we introduce a score correction mechanism based on a newly defined score difference. Our approach, which requires no network modifications, is validated across multiple models, datasets, and tasks, achieving state-of-the-art results.

1383. VibeCheck: Discover and Quantify Qualitative Differences in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29156 abstract： Large language models (LLMs) often exhibit subtle yet distinctive characteristics in their outputs that users intuitively recognize, but struggle to quantify. These "vibes" -- such as tone, formatting, or writing style -- influence user preferences, yet traditional evaluations focus primarily on the singular vibe of correctness.We introduce $\textbf{VibeCheck}$, a system for automatically comparing a pair of LLMs by discovering identifying traits of a model ("vibes") that are well-defined, differentiating, and user-aligned. VibeCheck iteratively discovers vibes from model outputs and then utilizes a panel of LLM judges to quantitatively measure the utility of each vibe. We validate that the vibes generated by VibeCheck align with those found in human discovery and run VibeCheck on pairwise preference data from real-world user conversations with Llama-3-70b vs GPT-4. VibeCheck reveals that Llama has a friendly, funny, and somewhat controversial vibe. These vibes predict model identity with 80% accuracy and human preference with 61% accuracy. Lastly, we run VibeCheck on a variety of models and tasks, including summarization, math, and captioning to provide insight into differences in model behavior. VibeCheck discovers vibes like Command X prefers to add concrete intros and conclusions when summarizing in comparison to TNGL, Llama-405b often overexplains its thought process on math problems compared to GPT-4o, and GPT-4 prefers to focus on the mood and emotions of the scene when captioning compared to Gemini-1.5-Flash.

1384. AutoCLIP: Auto-tuning Zero-Shot Classifiers for Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/31476 abstract： Classifiers built upon vision-language models such as CLIP have shown remarkable zero-shot performance across a broad range of image classification tasks. Prior work has studied different ways of automatically creating descriptor sets for every class based on prompt templates, ranging from manually engineered templates over templates obtained from a large language model to templates built from random words and characters. Up until now, deriving zero-shot classifiers from the respective encoded class descriptors has remained nearly unchanged, i.e., classify to the class that maximizes cosine similarity between its averaged encoded class descriptors and the image encoding. However, weighing all class descriptors equally can be suboptimal when certain descriptors match visual clues on a given image better than others. In this work, we propose AutoCLIP, a method for auto-tuning zero-shot classifiers. AutoCLIP tunes per-image weights to each prompt template at inference time, based on statistics of class descriptor-image similarities. AutoCLIP is fully unsupervised, has only a minor additional computation overhead, and can be easily implemented in few lines of code. We show that AutoCLIP outperforms baselines across a broad range of vision-language models, datasets, and prompt templates consistently and by up to 3 percent point accuracy.

1385. Improved Sampling Algorithms for Lévy-Itô Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29281 abstract： Lévy-Itô denoising diffusion models relying on isotropic α-stable noise instead of Gaussian distribution have recently been shown to improve performance of conventional diffusion models in image generation on imbalanced datasets while performing comparably in the standard settings. However, the stochastic algorithm of sampling from such models consists in solving the stochastic differential equation describing only an approximate inverse of the process of adding α-stable noise to data which may lead to suboptimal performance. In this paper, we derive a parametric family of stochastic differential equations whose solutions have the same marginal densities as those of the forward diffusion and show that the appropriate choice of the parameter values can improve quality of the generated images when the number of reverse diffusion steps is small. Also, we demonstrate that Lévy-Itô diffusion models are applicable to diverse domains and show that a well-trained text-to-speech Lévy-Itô model may have advantages over standard diffusion models on highly imbalanced datasets.

1386. Aioli: A Unified Optimization Framework for Language Model Data Mixing

链接：https://iclr.cc/virtual/2025/poster/28121 abstract： Language model performance depends on identifying the optimal mixture of data groups to train on (e.g., law, code, math). Prior work has proposed a diverse set of methods to efficiently learn mixture proportions, ranging from fitting regression models over training runs to dynamically updating proportions throughout training. Surprisingly, we find that no existing method consistently outperforms a simple stratified sampling baseline in terms of average test perplexity. To understand this inconsistency, we unify existing methods into a standard framework, showing they are equivalent to solving a common optimization problem: minimize average loss subject to a method-specific mixing law---an implicit assumption on the relationship between loss and mixture proportions. This framework suggests that measuring the fidelity of a method's mixing law can offer insights into its performance. Empirically, we find that existing methods set their mixing law parameters inaccurately, resulting in the inconsistent mixing performance we observe. Using this insight, we derive a new online method named Aioli, which directly estimates the mixing law parameters throughout training and uses them to dynamically adjust proportions. Empirically, Aioli outperforms stratified sampling on 6 out of 6 datasets by an average of 0.27 test perplexity points, whereas existing methods fail to consistently beat stratified sampling, doing up to 6.9 points worse. Moreover, in a practical setting where proportions are learned on shorter runs due to computational constraints, Aioli can dynamically adjust these proportions over the full training run, consistently improving performance over existing methods by up to 12.012 test perplexity points.

1387. ActionReasoningBench: Reasoning about Actions with and without Ramification Constraints

链接：https://iclr.cc/virtual/2025/poster/29872 abstract： Reasoning about Actions and Change (RAC) has historically played a pivotal role in solving foundational AI problems, such as the frame problem. It has driven advancements in AI fields, such as non-monotonic and commonsense reasoning. RAC remains crucial for AI systems that operate in dynamic environments, engage in interactive scenarios, or rely on commonsense reasoning. Despite substantial advances made by Large Language Models (LLMs) in various AI domains, their performance in RAC remains underexplored. To address this gap, we introduce a new diagnostic benchmark, $\textbf{ActionReasoningBench}$, which encompasses 8 domains and includes questions for up to 19 action sequences. This benchmark rigorously evaluates LLMs across six key RAC dimensions: $\textit{Fluent Tracking}$, $\textit{State Tracking}$, $\textit{Action Executability}$, $\textit{Effects of Actions}$, $\textit{Numerical RAC}$, and $\textit{Composite Questions}$. LLMs demonstrate average accuracy rates of 73.55%, 65.63%, 58.73%, and 62.38% on the former four dimensions, which are frequently discussed in RAC literature. However, the performance on the latter two dimensions, which introduce complex and novel reasoning questions, the average performance of LLMs is lowered to 33.16% and 51.19%, respectively, reflecting a 17.9% performance decline. We also introduce new ramification constraints to capture the indirect effects of actions, providing deeper insights into RAC challenges. Our evaluation of state-of-the-art LLMs, including both open-source and commercial models, reveals challenges across all RAC dimensions, particularly in handling ramifications, with GPT-4o failing to solve any question and o1-preview achieving a score of only 18.4%.

1388. Counterfactual Realizability

链接：https://iclr.cc/virtual/2025/poster/27947 abstract： It is commonly believed that, in a real-world environment, samples can only be drawn from observational and interventional distributions, corresponding to Layers 1 and 2 of the Pearl Causal Hierarchy. Layer 3, representing counterfactual distributions, is believed to be inaccessible by definition. However, Bareinboim, Forney, and Pearl (2015) introduced a procedure that allows an agent to sample directly from a counterfactual distribution, leaving open the question of what other counterfactual quantities can be estimated directly via physical experimentation. We resolve this by introducing a formal definition of realizability, the ability to draw samples from a distribution, and then developing a complete algorithm to determine whether an arbitrary counterfactual distribution is realizable given fundamental physical constraints, such as the inability to go back in time and subject the same unit to a different experimental condition. We illustrate the implications of this new framework for counterfactual data collection using motivating examples from causal fairness and causal reinforcement learning. While the baseline approach in these motivating settings typically follows an interventional or observational strategy, we show that a counterfactual strategy provably dominates both.

1389. MMD-Regularized Unbalanced Optimal Transport

链接：https://iclr.cc/virtual/2025/poster/31505 abstract： We study the unbalanced optimal transport (UOT) problem, where the marginal constraints are enforced using Maximum Mean Discrepancy (MMD) regularization. Our work is motivated by the observation that the literature on UOT is focused on regularization based on ϕ-divergence (e.g., KL divergence). Despite the popularity of MMD, its role as a regularizer in the context of UOT seems less understood. We begin by deriving a specific dual of MMD-regularized UOT (MMD-UOT), which helps us prove several useful properties. One interesting outcome of this duality result is that MMD-UOT induces novel metrics, which not only lift the ground metric like the Wasserstein but are also sample-wise efficient to estimate like the MMD. Further, for real-world applications involving non-discrete measures, we present an estimator for the transport plan that is supported only on the given (m) samples. Under certain conditions, we prove that the estimation error with this finitely-supported transport plan is also $\mathcal{O}(1/\sqrt{m})$. As far as we know, such error bounds that are free from the curse of dimensionality are not known for ϕ-divergence regularized UOT. Finally, we discuss how the proposed estimator can be computed efficiently using accelerated gradient descent. Our experiments show that MMD-UOT consistently outperforms popular baselines, including KL-regularized UOT and MMD, in diverse machine learning applications.

1390. Variance-Reducing Couplings for Random Features

链接：https://iclr.cc/virtual/2025/poster/28364 abstract： Random features (RFs) are a popular technique to scale up kernel methods in machine learning, replacing exact kernel evaluations with stochastic Monte Carlo estimates. They underpin models as diverse as efficient transformers (by approximating attention) to sparse spectrum Gaussian processes (by approximating the covariance function). Efficiency can be further improved by speeding up the convergence of these estimates: a variance reduction problem. We tackle this through the unifying lens of optimal transport, finding couplings to improve RFs defined on both Euclidean and discrete input spaces. They enjoy theoretical guarantees and sometimes provide strong downstream gains, including for scalable inference on graphs. We reach surprising conclusions about the benefits and limitations of variance reduction as a paradigm, showing that other properties of the coupling should be optimised for attention estimation in efficient transformers.

1391. Geometry of Neural Reinforcement Learning in Continuous State and Action Spaces

链接：https://iclr.cc/virtual/2025/poster/30635 abstract： Advances in reinforcement learning (RL) have led to its successful application in complex tasks with continuous state and action spaces. Despite these advances in practice, most theoretical work pertains to finite state and action spaces. We propose building a theoretical understanding of continuous state and action spaces by employing a geometric lens to understand the locally attained set of states. The set of all parametrised policies learnt through a semi-gradient based approach induce a set of attainable states in RL. We show that training dynamics of a two layer neural policy induce a low dimensional manifold of attainable states embedded in the high-dimensional nominal state space trained using an actor-critic algorithm. We prove that, under certain conditions, the dimensionality of this manifold is of the order of the dimensionality of the action space. This is the first result of its kind, linking the geometry of the state space to the dimensionality of the action space. We empirically corroborate this upper bound for four MuJoCo environments and also demonstrate the results in a toy environment with varying dimensionality. We also show the applicability of this theoretical result by introducing a local manifold learning layer to the policy and value function networks to improve the performance in control environments with very high degrees of freedom by changing one layer of the neural network to learn sparse representations.

1392. Differentiable and Learnable Wireless Simulation with Geometric Transformers

链接：https://iclr.cc/virtual/2025/poster/30689 abstract： Modelling the propagation of electromagnetic wireless signals is critical for designing modern communication systems. Wireless ray tracing simulators model signal propagation based on the 3D geometry and other scene parameters, but their accuracy is fundamentally limited by underlying modelling assumptions and correctness of parameters. In this work, we introduce Wi-GATr, a fully-learnable neural simulation surrogate designed to predict the channel observations based on scene primitives (e. g., surface mesh, antenna position and orientation). Recognizing the inherently geometric nature of these primitives, Wi-GATr leverages an equivariant Geometric Algebra Transformer that operates on a tokenizer specifically tailored for wireless simulation. We evaluate our approach on a range of tasks (i. e., signal strength and delay spread prediction, receiver localization, and geometry reconstruction) and find that Wi-GATr is accurate, fast, sample-efficient, and robust to symmetry-induced transformations. Remarkably, we find our results also translate well to the real world: Wi-GATr demonstrates more than 35% lower error than hybrid techniques, and 70% lower error than a calibrated wireless tracer.

1393. Optimal Strong Regret and Violation in Constrained MDPs via Policy Optimization

链接：https://iclr.cc/virtual/2025/poster/30755 abstract： We study online learning in constrained MDPs (CMDPs), focusing on the goal of attaining sublinear strong regret and strong cumulative constraint violation. Differently from their standard (weak) counterparts, these metrics do not allow negative terms to compensate positive ones, raising considerable additional challenges. Efroni et al. (2020) were the first to propose an algorithm with sublinear strong regret and strong violation, by exploiting linear programming. Thus, their algorithm is highly inefficient, leaving as an open problem achieving sublinear bounds by means of policy optimization methods, which are much more efficient in practice. Very recently, Muller et al. (2024) have partially addressed this problem by proposing a policy optimization method that allows to attain $\widetilde{\mathcal{O}}(T^{0.93})$ strong regret/violation. This still leaves open the question of whether optimal bounds are achievable by using an approach of this kind. We answer such a question affirmatively, by providing an efficient policy optimization algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ strong regret/violation. Our algorithm implements a primal-dual scheme that employs a state-of-the-art policy optimization approach for adversarial (unconstrained) MDPs as primal algorithm, and a UCB-like update for dual variables.

1394. Do vision models perceive objects like toddlers ?

链接：https://iclr.cc/virtual/2025/poster/31352 abstract： Despite recent advances in artificial vision systems, humans are still more data-efficient at learning strong visual representations. Psychophysical experiments suggest that toddlers develop fundamental visual properties between the ages of one and three, which affect their perceptual system for the rest of their life. They begin to recognize impoverished variants of daily objects, pay more attention to the shape of an object to categorize it, prefer objects in specific orientations and progressively generalize over the configural arrangement of objects' parts. This post examines whether these four visual properties also emerge in off-the-shelf machine learning (ML) vision models. We reproduce and complement previous studies by comparing toddlers and a large set of diverse pre-trained vision models for each visual property. This way, we unveil the interplay between these visual properties and highlight the main differences between ML models and toddlers.

1395. Open-World Reinforcement Learning over Long Short-Term Imagination

链接：https://iclr.cc/virtual/2025/poster/27879 abstract： Training visual reinforcement learning agents in a high-dimensional open world presents significant challenges. While various model-based methods have improved sample efficiency by learning interactive world models, these agents tend to be “short-sighted”, as they are typically trained on short snippets of imagined experiences. We argue that the primary challenge in open-world decision-making is improving the exploration efficiency across a vast state space, especially for tasks that demand consideration of long-horizon payoffs. In this paper, we present LS-Imagine, which extends the imagination horizon within a limited number of state transition steps, enabling the agent to explore behaviors that potentially lead to promising long-term feedback. The foundation of our approach is to build a $\textit{long short-term world model}$. To achieve this, we simulate goal-conditioned jumpy state transitions and compute corresponding affordance maps by zooming in on specific areas within single images. This facilitates the integration of direct long-term values into behavior learning. Our method demonstrates significant improvements over state-of-the-art techniques in MineDojo.

1396. Flash Inference: Near Linear Time Inference for Long Convolution Sequence Models and Beyond

链接：https://iclr.cc/virtual/2025/poster/29040 abstract： While transformers have been at the core of most recent advancements in sequence generative models, their computational cost remains quadratic in sequence length.Several subquadratic architectures have been proposed to address this computational issue. Some of them, including long convolution sequence models (LCSMs), such as Hyena, address this issue at training time but remain quadratic during inference. We propose a method for speeding up LCSMs' exact inference to quasilinear time, identify the key properties that make this possible, and propose a general framework that exploits these. Our approach, inspired by previous work on relaxed polynomial interpolation, is based on a tiling which helps decrease memory movement and share computation. It has the added benefit of allowing for almost complete parallelization across layers of the position-mixing part of the architecture. Empirically, we provide a proof of concept implementation for Hyena, which gets up to $7.8\times$ end-to-end improvement over standard inference by improving $110\times$ within the position-mixing part.

1397. Expected Return Symmetries

链接：https://iclr.cc/virtual/2025/poster/27862 abstract： Symmetry is an important inductive bias that can improve model robustness and generalization across many deep learning domains. In multi-agent settings, a priori known symmetries have been shown to address a fundamental coordination failure mode known as mutually incompatible symmetry breaking; e.g. in a game where two independent agents can choose to move "left" or "right", and where a reward of +1 or -1 is received when the agents choose the same action or different actions, respectively. However, the efficient and automatic discovery of environment symmetries, in particular for decentralized partially observable Markov decision processes, remains an open problem. Furthermore, environmental symmetry breaking constitutes only one type of coordination failure, which motivates the search for a more accessible and broader symmetry class. In this paper, we introduce such a broader group of previously unexplored symmetries, which we call expected return symmetries, which contains environment symmetries as a subgroup. We show that agents trained to be compatible under the group of expected return symmetries achieve better zero-shot coordination results than those using environment symmetries. As an additional benefit, our method makes minimal a priori assumptions about the structure of their environment and does not require access to ground truth symmetries.

1398. Fast training and sampling of Restricted Boltzmann Machines

链接：https://iclr.cc/virtual/2025/poster/31062 abstract： Restricted Boltzmann Machines (RBMs) are powerful tools for modeling complex systems and extracting insights from data, but their training is hindered by the slow mixing of Markov Chain Monte Carlo (MCMC) processes, especially with highly structured datasets. In this study, we build on recent theoretical advances in RBM training and focus on the stepwise encoding of data patterns into singular vectors of the coupling matrix, significantly reducing the cost of generating new samples and evaluating the quality of the model, as well as the training cost in highly clustered datasets. The learning process is analogous to the thermodynamic continuous phase transitions observed in ferromagnetic models, where new modes in the probability measure emerge in a continuous manner. We leverage the continuous transitions in the training process to define a smooth annealing trajectory that enables reliable and computationally efficient log-likelihood estimates. This approach enables online assessment during training and introduces a novel sampling strategy called Parallel Trajectory Tempering (PTT) that outperforms previously optimized MCMC methods.To mitigate the critical slowdown effect in the early stages of training, we propose a pre-training phase. In this phase, the principal components are encoded into a low-rank RBM through a convex optimization process, facilitating efficient static Monte Carlo sampling and accurate computation of the partition function.Our results demonstrate that this pre-training strategy allows RBMs to efficiently handle highly structured datasets where conventional methods fail. Additionally, our log-likelihood estimation outperforms computationally intensive approaches in controlled scenarios, while the PTT algorithm significantly accelerates MCMC processes compared to conventional methods.

1399. Spreading Out-of-Distribution Detection on Graphs

链接：https://iclr.cc/virtual/2025/poster/28324 abstract： Node-level out-of-distribution (OOD) detection on graphs has received significant attention from the machine learning community. However, previous approaches are evaluated using unrealistic benchmarks that consider only randomly selected OOD nodes, failing to reflect the interactions among nodes. In this paper, we introduce a new challenging task to model the interactions of OOD nodes in a graph, termed spreading OOD detection, where a newly emerged OOD node spreads its property to neighboring nodes. We curate realistic benchmarks by employing the epidemic spreading models that simulate the spreading of OOD nodes on the graph. We also showcase a ``Spreading COVID-19'' dataset to demonstrate the applicability of spreading OOD detection in real-world scenarios. Furthermore, to effectively detect spreading OOD samples under the proposed benchmark setup, we present a new approach called energy distribution-based detector (EDBD), which includes a novel energy-aggregation scheme. EDBD is designed to mitigate undesired mixing of OOD scores between in-distribution (ID) and OOD nodes. Our extensive experimental results demonstrate the superiority of our approach over state-of-the-art methods in both spreading OOD detection and conventional node-level OOD detection tasks across seven benchmark datasets. The source code is available at https://github.com/daehoum1/edbd.

1400. Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Community Retrieval

链接：https://iclr.cc/virtual/2025/poster/30752 abstract： Large language models (LLMs) have demonstrated significant potential in clinical decision support. Yet LLMs still suffer from hallucinations and lack fine-grained contextual medical knowledge, limiting their high-stake healthcare applications such as clinical diagnosis. Traditional retrieval-augmented generation (RAG) methods attempt to address these limitations but frequently retrieve sparse or irrelevant information, undermining prediction accuracy. We introduce KARE, a novel framework that integrates knowledge graph (KG) community-level retrieval with LLM reasoning to enhance healthcare predictions. KARE constructs a comprehensive multi-source KG by integrating biomedical databases, clinical literature, and LLM-generated insights, and organizes it using hierarchical graph community detection and summarization for precise and contextually relevant information retrieval. Our key innovations include: (1) a dense medical knowledge structuring approach enabling accurate retrieval of relevant information; (2) a dynamic knowledge retrieval mechanism that enriches patient contexts with focused, multi-faceted medical insights; and (3) a reasoning-enhanced prediction framework that leverages these enriched contexts to produce both accurate and interpretable clinical predictions. Extensive experiments demonstrate that KARE outperforms leading models by up to 10.8-15.0\% on MIMIC-III and 12.6-12.7\% on MIMIC-IV for mortality and readmission predictions. In addition to its impressive prediction accuracy, our framework leverages the reasoning capabilities of LLMs, enhancing the trustworthiness of clinical predictions.