1001. Weakly-Supervised Affordance Grounding Guided by Part-Level Semantic Priors

链接：https://iclr.cc/virtual/2025/poster/31269 abstract： In this work, we focus on the task of weakly supervised affordance grounding, where a model is trained to identify affordance regions on objects using human-object interaction images and egocentric object images without dense labels. Previous works are mostly built upon class activation maps, which are effective for semantic segmentation but may not be suitable for locating actions and functions. Leveraging recent advanced foundation models, we develop a supervised training pipeline based on pseudo labels. The pseudo labels are generated from an off-the-shelf part segmentation model, guided by a mapping from affordance to part names.Furthermore, we introduce three key enhancements to the baseline model: a label refining stage, a fine-grained feature alignment process, and a lightweight reasoning module. These techniques harness the semantic knowledge of static objects embedded in off-the-shelf foundation models to improve affordance learning, effectively bridging the gap between objects and actions.Extensive experiments demonstrate that the performance of the proposed model has achieved a breakthrough improvement over existing methods.

1002. RMP-SAM: Towards Real-Time Multi-Purpose Segment Anything

链接：https://iclr.cc/virtual/2025/poster/31184 abstract： Recent segmentation methods, which adopt large-scale data training and transformer architecture, aim to create one foundation model that can perform multiple tasks. However, most of these methods rely on heavy encoder and decoder frameworks, hindering their performance in real-time scenarios. To explore real-time segmentation, recent advancements primarily focus on semantic segmentation within specific environments, such as autonomous driving. However, they often overlook the generalization ability of these models across diverse scenarios. Therefore, to fill this gap, this work explores a novel real-time segmentation setting called real-time multi-purpose segmentation. It contains three fundamental sub-tasks: interactive segmentation, panoptic segmentation, and video instance segmentation. Unlike previous methods, which use a specific design for each task, we aim to use only a single end-to-end model to accomplish all these tasks in real-time. To meet real-time requirements and balance multi-task learning, we present a novel dynamic convolution-based method, Real-Time Multi-Purpose SAM (RMP-SAM). It contains an efficient encoder and an efficient decoupled adapter to perform prompt-driven decoding. Moreover, we further explore different training strategies and one new adapter design to boost co-training performance further. We benchmark several strong baselines by extending existing works to support our multi-purpose segmentation. Extensive experiments demonstrate that RMP-SAM is effective and generalizes well on proposed benchmarks and other specific semantic tasks. Our implementation of RMP-SAM achieves the optimal balance between accuracy and speed for these tasks. The code is released at \url{https://github.com/xushilin1/RAP-SAM}

1003. Strength Estimation and Human-Like Strength Adjustment in Games

链接：https://iclr.cc/virtual/2025/poster/30488 abstract： Strength estimation and adjustment are crucial in designing human-AI interactions, particularly in games where AI surpasses human players. This paper introduces a novel strength system, including a strength estimator (SE) and an SE-based Monte Carlo tree search, denoted as SE-MCTS, which predicts strengths from games and offers different playing strengths with human styles. The strength estimator calculates strength scores and predicts ranks from games without direct human interaction. SE-MCTS utilizes the strength scores in a Monte Carlo tree search to adjust playing strength and style. We first conduct experiments in Go, a challenging board game with a wide range of ranks. Our strength estimator significantly achieves over 80% accuracy in predicting ranks by observing 15 games only, whereas the previous method reached 49% accuracy for 100 games. For strength adjustment, SE-MCTS successfully adjusts to designated ranks while achieving a 51.33% accuracy in aligning to human actions, outperforming a previous state-of-the-art, with only 42.56% accuracy. To demonstrate the generality of our strength system, we further apply SE and SE-MCTS to chess and obtain consistent results. These results show a promising approach to strength estimation and adjustment, enhancing human-AI interactions in games. Our code is available at https://rlg.iis.sinica.edu.tw/papers/strength-estimator.

1004. LASeR: Towards Diversified and Generalizable Robot Design with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30803 abstract： Recent advances in Large Language Models (LLMs) have stimulated a significant paradigm shift in evolutionary optimization, where hand-crafted search heuristics are gradually replaced with LLMs serving as intelligent search operators. However, these studies still bear some notable limitations, including a challenge to balance exploitation with exploration, often leading to inferior solution diversity, as well as poor generalizability of problem solving across different task settings. These unsolved issues render the prowess of LLMs in robot design automation largely untapped. In this work, we present LASeR -- Large Language Model-Aided Evolutionary Search for Robot Design Automation. Leveraging a novel reflection mechanism termed DiRect, we elicit more knowledgeable exploratory behaviors from LLMs based on past search trajectories, reshaping the exploration-exploitation tradeoff with dual improvements in optimization efficiency and solution diversity. Additionally, with evolution fully grounded in task-related background information, we unprecedentedly uncover the inter-task reasoning capabilities of LLMs, facilitating generalizable design processes that effectively inspire zero-shot robot proposals for new applications. Our simulated experiments on voxel-based soft robots showcase distinct advantages of LASeR over competitive baselines. Code at https://github.com/WoodySJR/LASeR.

1005. OptionZero: Planning with Learned Options

链接：https://iclr.cc/virtual/2025/poster/31084 abstract： Planning with options -- a sequence of primitive actions -- has been shown effective in reinforcement learning within complex environments. Previous studies have focused on planning with predefined options or learned options through expert demonstration data.Inspired by MuZero, which learns superhuman heuristics without any human knowledge, we propose a novel approach, named OptionZero. OptionZero incorporates an option network into MuZero, providing autonomous discovery of options through self-play games. Furthermore, we modify the dynamics network to provide environment transitions when using options, allowing searching deeper under the same simulation constraints. Empirical experiments conducted in 26 Atari games demonstrate that OptionZero outperforms MuZero, achieving a 131.58% improvement in mean human-normalized score. Our behavior analysis shows that OptionZero not only learns options but also acquires strategic skills tailored to different game characteristics. Our findings show promising directions for discovering and using options in planning. Our code is available at https://rlg.iis.sinica.edu.tw/papers/optionzero.

1006. Towards Automated Knowledge Integration From Human-Interpretable Representations

链接：https://iclr.cc/virtual/2025/poster/29873 abstract： A significant challenge in machine learning, particularly in noisy and low-data environments, lies in effectively incorporating inductive biases to enhance data efficiency and robustness. Despite the success of informed machine learning methods, designing algorithms with explicit inductive biases remains largely a manual process. In this work, we explore how prior knowledge represented in its native formats, e.g. in natural language, can be integrated into machine learning models in an automated manner. Inspired by the learning to learn principles of meta-learning, we consider the approach of learning to integrate knowledge via conditional meta-learning, a paradigm we refer to as informed meta-learning. We introduce and motivate theoretically the principles of informed meta-learning enabling automated and controllable inductive bias selection. To illustrate our claims, we implement an instantiation of informed meta-learning--the Informed Neural Process, and empirically demonstrate the potential benefits and limitations of informed meta-learning in improving data efficiency and generalisation.

1007. Divergence-Regularized Discounted Aggregation: Equilibrium Finding in Multiplayer Partially Observable Stochastic Games

链接：https://iclr.cc/virtual/2025/poster/30072 abstract： This paper presents Divergence-Regularized Discounted Aggregation (DRDA), a multi-round learning system for solving partially observable stochastic games (POSGs). DRDA is based on action values and applicable to multiplayer POSGs, which can unify normal-form games (NFGs), extensive-form games (EFGs) with perfect recall, and Markov games (MGs). In each single round, DRDA can be viewed as a discounted variant of Follow the Regularized Leader (FTRL) under a general value function for POSGs. While previous studies on discounted FTRL have demonstrated its last-iterate convergence towards quantal response equilibrium (QRE) in NFGs, this paper extends the theoretical results to POSGs under divergence regularization and generalizes the QRE concept of Nash distribution. The linear last-iterate convergence of single-round DRDA to its rest point is proved under the assumption on the hypomonotonicity of the game. When the rest point is unique, it induces the unique Nash distribution defined in the POSG, which has a bounded deviation from Nash equilibrium (NE). Under multiple learning rounds, DRDA keeps replacing the base policy for divergence regularization with the policy at the rest point in the previous round. It is further proved that the limit point of multi-round DRDA must be an exact NE (rather than a QRE). In experiments, discrete-time DRDA can converge to NE at a near-exponential rate in (multiplayer) NFGs and outperform the existing baselines for EFGs, MGs, and typical POSGs.

1008. Risk-Sensitive Diffusion: Robustly Optimizing Diffusion Models with Noisy Samples

链接：https://iclr.cc/virtual/2025/poster/29132 abstract： Diffusion models are mainly studied on image data. However, non-image data (e.g., tabular data) are also prevalent in real applications and tend to be noisy due to some inevitable factors in the stage of data collection, degrading the generation quality of diffusion models. In this paper, we consider a novel problem setting where every collected sample is paired with a vector indicating the data quality: risk vector. This setting applies to many scenarios involving noisy data and we propose risk-sensitive SDE, a type of stochastic differential equation (SDE) parameterized by the risk vector, to address it. With some proper coefficients, risk-sensitive SDE can minimize the negative effect of noisy samples on the optimization of diffusion models. We conduct systematic studies for both Gaussian and non-Gaussian noise distributions, providing analytical forms of risk-sensitive SDE. To verify the effectiveness of our method, we have conducted extensive experiments on multiple tabular and time-series datasets, showing that risk-sensitive SDE permits a robust optimization of diffusion models with noisy samples and significantly outperforms previous baselines.

1009. Schur's Positive-Definite Network: Deep Learning in the SPD cone with structure

链接：https://iclr.cc/virtual/2025/poster/27939 abstract： Estimating matrices in the symmetric positive-definite (SPD) cone is of interest for many applications ranging from computer vision to graph learning. While there exist various convex optimization-based estimators, they remain limited in expressivity due to their model-based approach. The success of deep learning motivates the use of learning-based approaches to estimate SPD matrices with neural networks in a data-driven fashion. However, designing effective neural architectures for SPD learning is challenging, particularly when the task requiresadditional

structural constraints, such as element-wise sparsity. Current approaches either do not ensure that the output meets all desired properties or lack expressivity. In this paper, we introduce SpodNet, a novel and generic learning module that guarantees SPD outputs and supports additional structural constraints. Notably, it solves the challenging task of learning jointly SPD andsparse matrices. Our experiments illustrate the versatility and relevance of SpodNet layers for such applications.

1010. Empowering LLM Agents with Zero-Shot Optimal Decision-Making through Q-learning

链接：https://iclr.cc/virtual/2025/poster/32096 abstract： Large language models (LLMs) are trained on extensive text data to gain general comprehension capability. Current LLM agents leverage this ability to make zero- or few-shot decisions without reinforcement learning (RL) but fail in making optimal decisions, as LLMs inherently perform next-token prediction rather than maximizing rewards. In contrast, agents trained via RL could make optimal decisions but require extensive environmental interaction. In this work, we develop an algorithm that combines the zero-shot capabilities of LLMs with the optimal decision-making of RL, referred to as the Model-based LLM Agent with Q-Learning (MLAQ). MLAQ employs Q-learning to derive optimal policies from transitions within memory. However, unlike RL agents that collect data from environmental interactions, MLAQ constructs an imagination space fully based on LLM to perform imaginary interactions for deriving zero-shot policies. Our proposed UCB variant generates high-quality imaginary data through interactions with the LLM-based world model, balancing exploration and exploitation while ensuring a sub-linear regret bound. Additionally, MLAQ incorporates a mixed-examination mechanism to filter out incorrect data. We evaluate MLAQ in benchmarks that present significant challenges for existing LLM agents. Results show that MLAQ achieves a optimal rate of over 90\% in tasks where other methods struggle to succeed. Additional experiments are conducted to reach the conclusion that introducing model-based RL into LLM agents shows significant potential to improve optimal decision-making ability. Our interactive website is available at http://mlaq.site.

1011. Discovering Group Structures via Unitary Representation Learning

链接：https://iclr.cc/virtual/2025/poster/29498 abstract： Discovering group structures within data poses a fundamental challenge across diverse scientific domains. A key obstacle is the non-differentiable nature of group axioms, hindering their integration into deep learning frameworks. To address this, we introduce a novel differentiable approach leveraging the representation theory of finite groups. Our method features a unique network architecture that models interactions between group elements via matrix multiplication of their representations, along with a regularizer promoting the unitarity of these representations. The interplay between the network architecture and the unitarity condition implicitly encourages the emergence of valid group structures. Evaluations demonstrate our method's ability to accurately recover group operations and their unitary representations from partial observations, achieving significant improvements in sample efficiency and a $\times 1000$ speedup over the state of the art. This work lays the foundation for a promising new paradigm in automated algebraic structure discovery, with potential applications across various domains, including automatic symmetry discovery for geometric deep learning.

1012. What's New in My Data? Novelty Exploration via Contrastive Generation

链接：https://iclr.cc/virtual/2025/poster/30165 abstract： Fine-tuning is widely used to adapt language models for specific goals, often leveraging real-world data such as patient records, customer-service interactions, or web content in languages not covered in pre-training.These datasets are typically massive, noisy, and often confidential, making their direct inspection challenging.However, understanding them is essential for guiding model deployment and informing decisions about data cleaning or suppressing any harmful behaviors learned during fine-tuning.In this study, we introduce the task of novelty discovery through generation, which aims to identify novel domains of a fine-tuning dataset by generating examples that illustrate these properties.Our approach — Contrastive Generative Exploration (CGE) — assumes no direct access to the data but instead relies on a pre-trained model and the same model after fine-tuning.By contrasting the predictions of these two models, CGE can generate examples that highlight novel domains of the fine-tuning data.However, this simple approach may produce examples that are too similar to one another, failing to capture the full range of novel domains present in the dataset.We address this by introducing an iterative version of CGE, where the previously generated examples are used to update the pre-trained model, and this updated model is then contrasted with the fully fine-tuned model to generate the next example, promoting diversity in the generated outputs.Our experiments demonstrate the effectiveness of CGE in detecting novel domains, such as toxic language, as well as new natural and programming languages.Furthermore, we show that CGE remains effective even when models are fine-tuned using differential privacy techniques.

1013. Language Agents Meet Causality -- Bridging LLMs and Causal World Models

链接：https://iclr.cc/virtual/2025/poster/27748 abstract： Large Language Models (LLMs) have recently shown great promise in planning and reasoning applications. These tasks demand robust systems, which arguably require a causal understanding of the environment. While LLMs can acquire and reflect common sense causal knowledge from their pretraining data, this information is often incomplete, incorrect, or inapplicable to a specific environment. In contrast, causal representation learning (CRL) focuses on identifying the underlying causal structure within a given environment. We propose a framework that integrates CRLs

with LLMs to enable causally-aware reasoning and planning. This framework learns a causal world model, with causal variables linked to natural language expressions. This mapping provides LLMs with a flexible interface to process and generate descriptions of actions and states in text form. Effectively, the causal world model acts as a simulator that the LLM can query and interact with. We evaluate the framework on causal inference and planning tasks across temporal scales and environmental complexities. Our experiments demonstrate the effectiveness of the approach, with the causally-aware method outperforming LLM-based reasoners, especially for longer planning horizons.

1014. Revisit Large-Scale Image-Caption Data in Pre-training Multimodal Foundation Models

链接：https://iclr.cc/virtual/2025/poster/29536 abstract： Recent advancements in multimodal models highlight the value of rewritten captions for improving performance, yet key challenges remain. For example, while synthetic captions often provide superior quality and image-text alignment, it is not clear whether they can fully replace AltTexts: the role of synthetic captions and their interaction with original web-crawled AltTexts in pre-training is still not well understood. Moreover, different multimodal foundation models may have unique preferences for specific caption formats, but efforts to identify the optimal captions for each model remain limited. In this work, we propose a novel, controllable, and scalable captioning pipeline designed to generate diverse caption formats tailored to various multimodal models. By examining short synthetic captions (SSC) and descriptive synthetic captions (DSC) as case studies, we systematically explore their effects and interactions with AltTexts across models such as CLIP, multimodal LLMs, and diffusion models. Our findings reveal that a hybrid approach that keeps both synthetic captions and AltTexts can outperform the use of synthetic captions alone, improving both alignment and performance, with each model demonstrating preferences for particular caption formats. This comprehensive analysis provides valuable insights into optimizing captioning strategies, thereby advancing the pre-training of multimodal foundation models.

1015. Rethinking Visual Counterfactual Explanations Through Region Constraint

链接：https://iclr.cc/virtual/2025/poster/28798 abstract： Visual counterfactual explanations (VCEs) have recently gained immense popularity as a tool for clarifying the decision-making process of image classifiers. This trend is largely motivated by what these explanations promise to deliver -- indicate semantically meaningful factors that change the classifier's decision. However, we argue that current state-of-the-art approaches lack a crucial component -- the region constraint -- whose absence prevents from drawing explicit conclusions, and may even lead to faulty reasoning due to phenomenons like confirmation bias. To address the issue of previous methods, which modify images in a very entangled and widely dispersed manner, we propose region-constrained VCEs (RVCEs), which assume that only a predefined image region can be modified to influence the model's prediction. To effectively sample from this subclass of VCEs, we propose Region-Constrained Counterfactual Schrödinger Bridge (RCSB), an adaptation of a tractable subclass of Schrödinger Bridges to the problem of conditional inpainting, where the conditioning signal originates from the classifier of interest. In addition to setting a new state-of-the-art by a large margin, we extend RCSB to allow for exact counterfactual reasoning, where the predefined region contains only the factor of interest, and incorporating the user to actively interact with the RVCE by predefining the regions manually.

1016. IGL-Bench: Establishing the Comprehensive Benchmark for Imbalanced Graph Learning

链接：https://iclr.cc/virtual/2025/poster/27975 abstract： Deep graph learning has gained grand popularity over the past years due to its versatility and success in representing graph data across a wide range of domains. However, the pervasive issue of imbalanced graph data distributions, where certain parts exhibit disproportionally abundant data while others remain sparse, undermines the efficacy of conventional graph learning algorithms, leading to biased outcomes. To address this challenge, Imbalanced Graph Learning (IGL) has garnered substantial attention, enabling more balanced data distributions and better task performance. Despite the proliferation of IGL algorithms, the absence of consistent experimental protocols and fair performance comparisons pose a significant barrier to comprehending advancements in this field. To bridge this gap, we introduce IGL-Bench, a foundational comprehensive benchmark for imbalanced graph learning, embarking on 17 diverse graph datasets and 24 distinct IGL algorithms with uniform data processing and splitting strategies. Specifically, IGL-Bench systematically investigates state-of-the-art IGL algorithms in terms of effectiveness, robustness, and efficiency on node-level and graph-level tasks, with the scope of class-imbalance and topology-imbalance. Extensive experiments demonstrate the potential benefits of IGL algorithms on various imbalanced conditions, offering insights and opportunities in the IGL field. Further, we have developed an open-sourced and unified package to facilitate reproducible evaluation and inspire further innovative research, available at: https://github.com/RingBDStack/IGL-Bench.

1017. Beyond Random Masking: When Dropout meets Graph Convolutional Networks

链接：https://iclr.cc/virtual/2025/poster/29723 abstract： Graph Convolutional Networks (GCNs) have emerged as powerful tools for learning on graph-structured data, yet the behavior of dropout in these models remains poorly understood. This paper presents a comprehensive theoretical analysis of dropout in GCNs, revealing that its primary role differs fundamentally from standard neural networks - preventing oversmoothing rather than co-adaptation. We demonstrate that dropout in GCNs creates

dimension-specific stochastic sub-graphs, leading to a form of structural regularization not present in standard neural networks. Our analysis shows that dropout effects are inherently degree-dependent, resulting in adaptive regularization that considers the topological importance of nodes. We provide new insights into dropout's role in mitigating oversmoothing and derive novel generalization bounds that account for graph-specific dropout effects. Furthermore, we analyze the synergistic interaction between dropout and batch normalization in GCNs, uncovering a mechanism that enhances overall regularization. Our theoretical findings are validated through extensive experiments on both node-level and graph-level tasks across 14 datasets. Notably, GCN with dropout and batch normalization outperforms state-of-the-art methods on several benchmarks, demonstrating the practical impact of our theoretical insights.

1018. Not-So-Optimal Transport Flows for 3D Point Cloud Generation

链接：https://iclr.cc/virtual/2025/poster/30911 abstract： Learning generative models of 3D point clouds is one of the fundamental problems in 3D generative learning. One of the key properties of point clouds is their permutation invariance, i.e., changing the order of points in a point cloud does not change the shape they represent. In this paper, we analyze the recently proposed equivariant OT flows that learn permutation invariant generative models for point-based molecular data and we show that these models scale poorly on large point clouds. Also, we observe learning (equivariant) OT flows is generally challenging since straightening flow trajectories makes the learned flow model complex at the beginning of the trajectory. To remedy these, we propose not-so-optimal transport flow models that obtain an approximate OT by an offline OT precomputation, enabling an efficient construction of OT pairs for training. During training, we can additionally construct a hybrid coupling by combining our approximate OT and independent coupling to make the target flow models easier to learn. In an extensive empirical study, we show that our proposed model outperforms prior diffusion- and flow -based approaches on a wide range of unconditional generation and shape completion on the ShapeNet benchmark.

1019. Node Identifiers: Compact, Discrete Representations for Efficient Graph Learning

链接：https://iclr.cc/virtual/2025/poster/28071 abstract： We present a novel end-to-end framework that generates highly compact (typically 6-15 dimensions), discrete (int4 type), and interpretable node representations—termed node identifiers (node IDs)—to tackle inference challenges on large-scale graphs. By employing vector quantization, we compress continuous node embeddings from multiple layers of a Graph Neural Network (GNN) into discrete codes, applicable under both self-supervised and supervised learning paradigms. These node IDs capture high-level abstractions of graph data and offer interpretability that traditional GNN embeddings lack. Extensive experiments on 34 datasets, encompassing node classification, graph classification, link prediction, and attributed graph clustering tasks, demonstrate that the generated node IDs significantly enhance speed and memory efficiency while achieving competitive performance compared to current state-of-the-art methods. Our source code is available at https://github.com/LUOyk1999/NodeID.

1020. MoE++: Accelerating Mixture-of-Experts Methods with Zero-Computation Experts

链接：https://iclr.cc/virtual/2025/poster/28078 abstract： In this work, we aim to simultaneously enhance the effectiveness and efficiency of Mixture-of-Experts (MoE) methods. To achieve this, we propose MoE++, a general and heterogeneous MoE framework that integrates both Feed-Forward Network (FFN) and zero-computation experts. Specifically, we introduce three types of zero-computation experts: the zero expert, copy expert, and constant expert, which correspond to discard, skip, and replace operations, respectively. This design offers three key advantages: (i) **Low Computing Overhead**: Unlike the uniform mixing mechanism for all tokens within vanilla MoE, MoE++ allows each token to engage with a dynamic number of FFNs, be adjusted by constant vectors, or even skip the MoE layer entirely. (ii) **High Performance**: By enabling simple tokens to utilize fewer FFN experts, MoE++ allows more experts to focus on challenging tokens, thereby unlocking greater performance potential than vanilla MoE. (iii) **Deployment Friendly**: Given that zero-computation experts have negligible parameters, we can deploy all zero-computation experts on each GPU, eliminating the significant communication overhead and expert load imbalance associated with FFN experts distributed across different GPUs. Moreover, we leverage gating residuals, enabling each token to consider the pathway taken in the previous layer when selecting the appropriate experts. Extensive experimental results demonstrate that MoE++ achieves better performance while delivering 1.1$\sim$2.1$\times$ expert forward throughput compared to a vanilla MoE model of the same size, which lays a solid foundation for developing advanced and efficient MoE-related models.

1021. Pacmann: Efficient Private Approximate Nearest Neighbor Search

链接：https://iclr.cc/virtual/2025/poster/27729 abstract： We propose a new private Approximate Nearest Neighbor (ANN) search schemenamed Pacmannthat allows a client to perform ANN searchin a vector database without revealing the query vector to the server.Unlike prior constructions that run encrypted search on the server side,Pacmann carefully offloads limited computation and storage to the client,no longer requiring computationally-intensive cryptographic techniques.Specifically, clients run a graph-based ANN search, where in each hop on the graph, the client privately retrieves local graph information from the server. To make this efficient, we combine two ideas: (1) we adapt a leading graph-based ANN search algorithm to be compatible with private information retrieval (PIR) for subgraph retrieval;(2) we use a recent class of PIR schemes that trade offline preprocessing for online computational efficiency. Pacmann achieves significantly better search quality thanthe state-of-

the-art private ANN search schemes,showing up to 2.5$\times$ better search accuracy on real-world datasets than prior work andreaching 90\% quality of a state-of-the-art non-private ANN algorithm.Moreover on large datasets with up to 100 million vectors,Pacmann shows better scalability than prior private ANN schemeswith up to 62\% reduction in computation timeand 22\% reduction in overall latency.

1022. PiCO: Peer Review in LLMs based on Consistency Optimization

链接：https://iclr.cc/virtual/2025/poster/28108 abstract： Existing large language models (LLMs) evaluation methods typically focus on testing the performance on some closed-environment and domain-specific benchmarks with human annotations. In this paper, we explore a novel unsupervised evaluation direction, utilizing peer-review mechanisms to measure LLMs automatically without any human feedback. In this setting, both open-source and closed-source LLMs lie in the same environment, capable of answering unlabeled questions and evaluating each other, where each LLM's response score is jointly determined by other anonymous ones. During this process, we found that those answers that are more recognized by other ``reviewers'' (models) usually come from LLMs with stronger abilities, while these models can also evaluate others' answers more accurately. We formalize it as a consistency assumption, i.e., the ability and score of the model usually have consistency. We exploit this to optimize each model's confidence, thereby re-ranking the LLMs to be closer to human rankings. We perform experiments on multiple datasets with standard rank-based metrics, validating the effectiveness of the proposed approach.

1023. RRM: Robust Reward Model Training Mitigates Reward Hacking

链接：https://iclr.cc/virtual/2025/poster/30783 abstract： Reward models (RMs) play a pivotal role in aligning large language models (LLMs) with human preferences. However, traditional RM training, which relies on response pairs tied to specific prompts, struggles to disentangle prompt-driven preferences from prompt-independent artifacts, such as response length and format. In this work, we expose a fundamental limitation of current RM training methods, where RMs fail to effectively distinguish between contextual signals and irrelevant artifacts when determining preferences. To address this, we introduce a causal framework that learns preferences independent of these artifacts and propose a novel data augmentation technique designed to eliminate them. Extensive experiments show that our approach successfully filters out undesirable artifacts, yielding a more robust reward model (RRM). Our RRM improves the performance of a pairwise reward model trained on Gemma-2-9b-it, on Reward-Bench, increasing accuracy from 80.61% to 84.15%. Additionally, we train two DPO policies using both the RM and RRM, demonstrating that the RRM significantly enhances DPO-aligned policies, improving MT-Bench scores from 7.27 to 8.31 and length-controlled win-rates in AlpacaEval-2 from 33.46% to 52.49%.

1024. MMEgo: Towards Building Egocentric Multimodal LLMs for Video QA

链接：https://iclr.cc/virtual/2025/poster/30907 abstract： This research aims to comprehensively explore building a multimodal foundation model for egocentric video understanding.To achieve this goal, we work on three fronts. First, as there is a lack of QA data for egocentric video understanding, we automatically generate 7M high-quality QA samples for egocentric videos ranging from 30 seconds to one hour long in Ego4D based on human-annotated data.This is one of the largest egocentric QA datasets.Second, we contribute a challenging egocentric QA benchmark with 629 videos and 7,026 questions to evaluate the models' ability in recognizing and memorizing visual details across videos of varying lengths. We introduce a new de-biasing evaluation method to help mitigate the unavoidable language bias present in the models being evaluated.Third, we propose a specialized multimodal architecture featuring a novel ``Memory Pointer Prompting'' mechanism. This design includes a global glimpse step to gain an overarching understanding of the entire video and identify key visual information, followed by a fallback step that utilizes the key visual information to generate responses. This enables the model to more effectively comprehend extended video content.With the data, benchmark, and model, we build MM-Ego, an egocentric multimodal LLM that shows powerful performance on egocentric video understanding.

1025. Can We Ignore Labels in Out of Distribution Detection?

链接：https://iclr.cc/virtual/2025/poster/28870 abstract： Out-of-distribution (OOD) detection methods have recently become more prominent, serving as a core element in safety-critical autonomous systems. One major purpose of OOD detection is to reject invalid inputs that could lead to unpredictable errors and compromise safety. Due to the cost of labeled data, recent works have investigated the feasibility of self-supervised learning (SSL) OOD detection, unlabled OOD detection, and zero shot OOD detection. In this work, we identify a set of conditions for a theoretical guarantee of failure in unlabeled OOD detection algorithms from an information-theoretic perspective. These conditions are present in all OOD tasks dealing with real world data: I) we provide theoretical proof of unlabeled OOD detection failure when there exists zero mutual information between the learning objective and the in-distribution labels, a.k.a. 'label blindness', II) we define a new OOD task – Adjacent OOD detection – that tests for label blindness and accounts for a previously ignored safety gap in all OOD detection benchmarks, and III) we perform experiments demonstrating that existing unlabeled OOD methods fail under conditions suggested by our label blindness theory and analyze the implications for future research in unlabeled OOD methods.

1026. MoDGS: Dynamic Gaussian Splatting from Casually-captured Monocular Videos with Depth Priors

链接：https://iclr.cc/virtual/2025/poster/31112 abstract： In this paper, we propose MoDGS, a new pipeline to render novel-view

images in dynamic scenes using only casually captured monocular videos. Previous monocular dynamic NeRF or Gaussian Splatting methods strongly rely on the rapid movement of input cameras to construct multiview consistency but fail to reconstruct dynamic scenes on casually captured input videos whose cameras are static or move slowly. To address this challenging task, MoDGS adopts recent single-view depth estimation methods to guide the learning of the dynamic scene. Then, a novel 3D-aware initialization method is proposed to learn a reasonable deformation field and a new robust depth loss is proposed to guide the learning of dynamic scene geometry. Comprehensive experiments demonstrate that MoDGS is able to render high-quality novel view images of dynamic scenes from just a casually captured monocular video, which outperforms baseline methods by a significant margin. Project page: https://MoDGS.github.io

1027. A Non-Contrastive Learning Framework for Sequential Recommendation with Preference-Preserving Profile Generation

链接：https://iclr.cc/virtual/2025/poster/30042 abstract： Contrastive Learning (CL) proves to be effective for learning generalizable user representations in Sequential Recommendation (SR), but it suffers from high computational costs due to its reliance on negative samples. To overcome this limitation, we propose the first Non-Contrastive Learning (NCL) framework for SR, which eliminates computational overhead of identifying and generating negativesamples. However, without negative samples, it is challenging to learn uniform representations from only positive samples, which is prone to representation collapse. Furthermore, the alignment of the learned representations may be substantially compromised because existing ad-hoc augmentations can produce positive samples that have inconsistent user preferences. To tackle these challenges, we design a novel preference-preserving profile generation method to produce high-quality positive samples for non-contrastive training. Inspired by differential privacy, our approach creates augmented user profiles that exhibit high diversity while provably retaining consistent user preferences. With larger diversity and consistency of the positive samples, our NCL framework significantly enhances the alignment and uniformity of the learned representations, which contributes to better generalization. The experimental results on various benchmark datasets and model architectures demonstrate the effectiveness of the proposed method. Finally, our investigations reveal that both uniformity and alignment play a vital role in improving generalization for SR. Interestingly, in our data-sparse setting, alignment is usually more important than uniformity.

1028. The Same but Different: Structural Similarities and Differences in Multilingual Language Modeling

链接：https://iclr.cc/virtual/2025/poster/29893 abstract： We employ new tools from mechanistic interpretability to ask whether the internal structure of large language models (LLMs) shows correspondence to the linguistic structures which underlie the languages on which they are trained. In particular, we ask (1) when two languages employ the same morphosyntactic processes, do LLMs handle them using shared internal circuitry? and (2) when two languages require different morphosyntactic processes, do LLMs handle them using different internal circuitry? In a focused case study on English and Chinese multilingual and monolingual models, we analyze the internal circuitry involved in two tasks. We find evidence that models employ the same circuit to handle the same syntactic process independently of the language in which it occurs, and that this is the case even for monolingual models trained completely independently. Moreover, we show that multilingual models employ language-specific components (attention heads and feed-forward networks) when needed to handle linguistic processes (e.g., morphological marking) that only exist in some languages. Together, our results are revealing about how LLMs trade off between exploiting common structures and preserving linguistic differences when tasked with modeling multiple languages simultaneously, opening the door for future work in this direction.

1029. Training-Free Dataset Pruning for Instance Segmentation

链接：https://iclr.cc/virtual/2025/poster/28161 abstract： Existing dataset pruning techniques primarily focus on classification tasks, limiting their applicability to more complex and practical tasks like instance segmentation. Instance segmentation presents three key challenges: pixel-level annotations, instance area variations, and class imbalances, which significantly complicate dataset pruning efforts. Directly adapting existing classification-based pruning methods proves ineffective due to their reliance on time-consuming model training process. To address this, we propose a novel **T**raining-**F**ree **D**ataset **P**runing (**TFDP**) method for instance segmentation. Specifically, we leverage shape and class information from image annotations to design a Shape Complexity Score (SCS), refining it into a Scale-Invariant (SI-SCS) and Class-Balanced (CB-SCS) versions to address instance area variations and class imbalances, all without requiring model training. We achieve state-of-the-art results on VOC 2012, Cityscapes, and MS COCO datasets, generalizing well across CNN and Transformer architectures. Remarkably, our approach accelerates the pruning process by an average of **1349$\times$** on COCO compared to the adapted baselines.

1030. SEMDICE: Off-policy State Entropy Maximization via Stationary Distribution Correction Estimation

链接：https://iclr.cc/virtual/2025/poster/28205 abstract： In the unsupervised pre-training for reinforcement learning, the agent aims to learn a prior policy for downstream tasks without relying on task-specific reward functions. We focus on state entropy maximization (SEM), where the goal is to learn a policy that maximizes the entropy of the state's stationary distribution. In this paper, we introduce SEMDICE, a principled off-policy algorithm that computes an SEM policy from an arbitrary off-policy dataset, which optimizes the policy directly within the space of stationary distributions. SEMDICE computes a single, stationary

Markov state-entropy-maximizing policy from an arbitrary off-policy dataset. Experimental results demonstrate that SEMDICE outperforms baseline algorithms in maximizing state entropy while achieving the best adaptation efficiency for downstream tasks among SEM-based unsupervised RL pre-training methods.

1031. SpaceGNN: Multi-Space Graph Neural Network for Node Anomaly Detection with Extremely Limited Labels

链接：https://iclr.cc/virtual/2025/poster/29569 abstract： Node Anomaly Detection (NAD) has gained significant attention in the deep learning community due to its diverse applications in real-world scenarios. Existing NAD methods primarily embed graphs within a single Euclidean space, while overlooking the potential of non-Euclidean spaces. Besides, to address the prevalent issue of limited supervision in real NAD tasks, previous methods tend to leverage synthetic data to collect auxiliary information, which is not an effective solution as shown in our experiments.To overcome these challenges, we introduce a novel SpaceGNN model designed for NAD tasks with extremely limited labels. Specifically, we provide deeper insights into a task-relevant framework by empirically analyzing the benefits of different spaces for node representations, based on which, we design a Learnable Space Projection function that effectively encodes nodes into suitable spaces.Besides, we introduce the concept of weighted homogeneity, which we empirically and theoretically validate as an effective coefficient during information propagation. This concept inspires the design of the Distance Aware Propagation module. Furthermore, we propose the Multiple Space Ensemble module, which extracts comprehensive information for NAD under conditions of extremely limited supervision. Our findings indicate that this module is more beneficial than data augmentation techniques for NAD. Extensive experiments conducted on 9 real datasets confirm the superiority of SpaceGNN, which outperforms the best rival by an average of 8.55% in AUC and 4.31% in F1 scores. Our code is available at https://github.com/xydong127/SpaceGNN.

1032. ADePT: Adaptive Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/28848 abstract： Prompt Tuning (PT) enables the adaptation of Pre-trained Large Language Models (PLMs) to downstream tasks by optimizing a small amount of soft virtual tokens, which are prepended to the input token embeddings. Recently, Decomposed Prompt Tuning (DePT) has demonstrated superior adaptation capabilities by decomposing the soft prompt into a shorter soft prompt and a pair of low-rank matrices. The product of the pair of low-rank matrices is added to the input token embeddings to offset them. Additionally, DePT achieves faster inference compared to PT due to the shorter soft prompt. However, in this paper, we find that the position-based token embedding offsets of DePT restricts its ability to generalize across diverse model inputs, and that the shared embedding offsets across many token embeddings result in sub-optimization. To tackle these issues, we introduce \textbf{A}daptive \textbf{De}composed \textbf{P}rompt \textbf{T}uning (ADePT), which is composed of a short soft prompt and a shallow token-shared feed-forward neural network. ADePT utilizes the token-shared feed-forward neural network to learn the embedding offsets for each token, enabling adaptive embedding offsets that vary according to the model input and better optimization of token embedding offsets. This enables ADePT to achieve superior adaptation performance without requiring more inference time or additional trainable parameters compared to vanilla PT and its variants. In comprehensive experiments across 23 natural language processing tasks and 4 typical PLMs of different scales, ADePT consistently surpasses the leading parameter-efficient fine-tuning methods, and even outperforms the full fine-tuning in certain scenarios. We also provide a theoretical analysis towards ADePT. Code is available at https://github.com/HungerPWAY/ADePT.

1033. Boosting Neural Combinatorial Optimization for Large-Scale Vehicle Routing Problems

链接：https://iclr.cc/virtual/2025/poster/29530 abstract： Neural Combinatorial Optimization (NCO) methods have exhibited promising performance in solving Vehicle Routing Problems (VRPs). However, most NCO methods rely on the conventional self-attention mechanism that induces excessive computational complexity, thereby struggling to contend with large-scale VRPs and hindering their practical applicability. In this paper, we propose a lightweight cross-attention mechanism with linear complexity, by which a Transformer network is developed to learn efficient and favorable solutions for large-scale VRPs. We also propose a Self-Improved Training (SIT) algorithm that enables direct model training on large-scale VRP instances, bypassing extensive computational overhead for attaining labels. By iterating solution reconstruction, the Transformer network itself can generate improved partial solutions as pseudo-labels to guide the model training. Experimental results on the Travelling Salesman Problem (TSP) and the Capacitated Vehicle Routing Problem (CVRP) with up to 100K nodes indicate that our method consistently achieves superior performance for synthetic and real-world benchmarks, significantly boosting the scalability of NCO methods.

1034. Certifying Language Model Robustness with Fuzzed Randomized Smoothing: An Efficient Defense Against Backdoor Attacks

链接：https://iclr.cc/virtual/2025/poster/29472 abstract： The widespread deployment of pre-trained language models (PLMs) has exposed them to textual backdoor attacks, particularly those planted during the pre-training stage. These attacks pose significant risks to high-reliability applications, as they can stealthily affect multiple downstream tasks. While certifying robustness against such threats is crucial, existing defenses struggle with the high-dimensional, interdependent nature of textual

data and the lack of access to original poisoned pre-training data. To address these challenges, we introduce Fuzzed Randomized Smoothing (FRS), a novel approach for efficiently certifying language model robustness against backdoor attacks. FRS integrates software robustness certification techniques with biphased model parameter smoothing, employing Monte Carlo tree search for proactive fuzzing to identify vulnerable textual segments within the Damerau-Levenshtein space. This allows for targeted and efficient text randomization, while eliminating the need for access to poisoned training data during model smoothing. Our theoretical analysis demonstrates that FRS achieves a broader certified robustness radius compared to existing methods. Extensive experiments across various datasets, model configurations, and attack strategies validate FRS's superiority in terms of defense efficiency, accuracy, and robustness.

1035. A Single Goal is All You Need: Skills and Exploration Emerge from Contrastive RL without Rewards, Demonstrations, or Subgoals

链接：https://iclr.cc/virtual/2025/poster/27798 abstract： In this paper, we present empirical evidence of skills and directed exploration emerging from a simple RL algorithm long before any successful trials are observed. For example, in a manipulation task, the agent is given a single observation of the goal state (see Fig. 1) and learns skills, first for moving its end-effector, then for pushing the block, and finally for picking up and placing the block. These skills emerge before the agent has ever successfully placed the block at the goal location and without the aid of any reward functions, demonstrations, or manually-specified distance metrics. Once the agent has learned to reach the goal state reliably, exploration is reduced. Implementing our method involves a simple modification of prior work and does not require density estimates, ensembles, or any additional hyperparameters. Intuitively, the proposed method seems like it should be terrible at exploration, and we lack a clear theoretical understanding of why it works so effectively, though our experiments provide some hints.

1036. Task Descriptors Help Transformers Learn Linear Models In-Context

链接：https://iclr.cc/virtual/2025/poster/28511 abstract： Large language models (LLMs) exhibit strong in-context learning (ICL) ability, which allows the model to make predictions on new examples based on the given prompt. Recently, a line of research (Von Oswald et al., 2023; Aky¨urek et al., 2023; Ahn et al., 2023; Mahankali et al., 2023; Zhang et al., 2024) considered ICL for a simple linear regression setting and showed that the forward pass of Transformers is simulating some variants of gradient descent (GD) algorithms on the in-context examples. In practice, the input prompt usually contains a task descriptor in addition to in-context examples. We investigate how the task description helps ICL in the linear regression setting. Consider a simple setting where the task descriptor describes the mean of input in linear regression. Our results show that gradient flow converges to a global minimum for a linear Transformer. At the global minimum, the Transformer learns to use the task descriptor effectively to improve its performance. Empirically, we verify our results by showing that the weights converge to the predicted global minimum and Transformers indeed perform better with task descriptors.

1037. Fast Direct: Query-Efficient Online Black-box Guidance for Diffusion-model Target Generation

链接：https://iclr.cc/virtual/2025/poster/29800 abstract： Guided diffusion-model generation is a promising direction for customizing the generation process of a pre-trained diffusion model to address specific downstream tasks. Existing guided diffusion models either rely on training the guidance model with pre-collected datasets or require the objective functions to be differentiable. However, for most real-world tasks, offline datasets are often unavailable, and their objective functions are often not differentiable, such as image generation with human preferences, molecular generation for drug discovery, and material design. Thus, we need an **online** algorithm capable of collecting data during runtime and supporting a **black-box** objective function. Moreover, the **query efficiency** of the algorithm is also critical because the objective evaluation of the query is often expensive in real-world scenarios. In this work, we propose a novel and simple algorithm, **Fast Direct**, for query-efficient online black-box target generation. Our Fast Direct builds a pseudo-target on the data manifold to update the noise sequence of the diffusion model with a universal direction, which is promising to perform query-efficient guided generation. Extensive experiments on twelve high-resolution ($\small {1024 \times 1024}$) image target generation tasks and six 3D-molecule target generation tasks show $\textbf{6}\times$ up to $\textbf{10}\times$ query efficiency improvement and $\textbf{11}\times$ up to $\textbf{44}\times$ query efficiency improvement, respectively.

1038. Boosting the visual interpretability of CLIP via adversarial fine-tuning

链接：https://iclr.cc/virtual/2025/poster/28566 abstract： CLIP has achieved great success in visual representation learning and is becoming an important plug-in component for many large multi-modal models like LLaVA and DALL-E. However, the lack of interpretability caused by the intricate image encoder architecture and training process restricts its wider use in high-stake decision making applications. In this work, we propose an unsupervised adversarial fine-tuning (AFT) with norm-regularization to enhance the visual interpretability of CLIP. We provide theoretical analysis showing that AFT has implicit regularization that enforces the image encoder to encode the input features sparsely, directing the network's focus towards meaningful features. Evaluations by both feature attribution techniques and network dissection offer convincing evidence that the visual interpretability of CLIP has significant improvements. With AFT, the image encoder prioritizes pertinent input features, and the neuron within the encoder exhibits better alignment with human-understandable concepts. Moreover, these effects are generalizable to out-of-distribution datasets and can be transferred to downstream tasks. Additionally, AFT enhances the visual interpretability of derived large vision-language models that incorporate the pre-trained CLIP an integral component. The code of this paper is

available at the CLIP_AFT GitHub repository.

1039. LLMs Can Plan Only If We Tell Them

链接：https://iclr.cc/virtual/2025/poster/30078 abstract： Large language models (LLMs) have demonstrated significant capabilities in natural language processing and reasoning, yet their effectiveness in autonomous planning has been under debate. While existing studies have utilized LLMs with external feedback mechanisms or in controlled environments for planning, these approaches often involve substantial computational and development resources due to the requirement for careful design and iterative backprompting. Moreover, even the most advanced LLMs like GPT-4 struggle to match human performance on standard planning benchmarks, such as the Blocksworld, without additional support. This paper investigates whether LLMs can independently generate long-horizon plans that rival human baselines. Our novel enhancements to Algorithm-of-Thoughts (AoT), which we dub AoT+, help achieve state-of-the-art results in planning benchmarks out-competing prior methods and human baselines all autonomously.

1040. Topological Schrödinger Bridge Matching

链接：https://iclr.cc/virtual/2025/poster/29332 abstract： Given two boundary distributions, the \emph{Schrödinger Bridge} (SB) problem seeks the “most likely” random evolution between them with respect to a reference process. It has revealed rich connections to recent machine learning methods for generative modeling and distribution matching. While these methods perform well in Euclidean domains, they are not directly applicable to topological domains such as graphs and simplicial complexes, which are crucial for data defined over network entities, such as node signals and edge flows. In this work, we propose the \emph{Topological Schrödinger Bridge problem} ($\mathcal{T}$SBP) for matching signal distributions on a topological domain. We set the reference process to follow some linear tractable \emph{topology-aware} stochastic dynamics such as topological heat diffusion. For the case of Gaussian boundary distributions, we derive a \emph{closed-form} topological SB ($\mathcal{T}$SB) in terms of its time-marginal and stochastic differential. In the general case, leveraging the well-known result, we show that the optimal process follows the forward-backward topological dynamics governed by some unknowns. Building on these results, we develop $\mathcal{T}$SB-based models for matching topological signals by parameterizing the unknowns in the optimal process as \emph{(topological) neural networks} and learning them through \emph{likelihood training}. We validate the theoretical results and demonstrate the practical applications of $\mathcal{T}$SB-based models on both synthetic and real-world networks, emphasizing the role of topology. Additionally, we discuss the connections of $\mathcal{T}$SB-based models to other emerging models, and outline future directions for topological signal matching.

1041. Extending Mercer's expansion to indefinite and asymmetric kernels

链接：https://iclr.cc/virtual/2025/poster/28631 abstract： Mercer's expansion and Mercer's theorem are cornerstone results in kernel theory. While the classical Mercer's theorem only considers continuous symmetric positive definite kernels, analogous expansions are effective in practice for indefinite and asymmetric kernels. In this paper we extend Mercer's expansion to continuous kernels, providing a rigorous theoretical underpinning for indefinite and asymmetric kernels. We begin by demonstrating that Mercer's expansion may not be pointwise convergent for continuous indefinite kernels, before proving that the expansion of continuous kernels with bounded variation uniformly in each variable separably converges pointwise almost everywhere, almost uniformly, and unconditionally almost everywhere. We also describe an algorithm for computing Mercer's expansion for general kernels and give new decay bounds on its terms.

1042. FreDF: Learning to Forecast in the Frequency Domain

链接：https://iclr.cc/virtual/2025/poster/31031 abstract： Time series modeling presents unique challenges due to autocorrelation in both historical data and future sequences. While current research predominantly addresses autocorrelation within historical data, the correlations among future labels are often overlooked. Specifically, modern forecasting models primarily adhere to the Direct Forecast (DF) paradigm, generating multi-step forecasts independently and disregarding label correlations over time. In this work, we demonstrate that the learning objective of DF is biased in the presence of label correlation. To address this issue, we propose the Frequency-enhanced Direct Forecast (FreDF), which mitigates label correlation by learning to forecast in the frequency domain, thereby reducing estimation bias. Our experiments show that FreDF significantly outperforms existing state-of-the-art methods and is compatible with a variety of forecast models. Code is available at https://github.com/Master-PLC/FreDF.

1043. Differentially Private Federated Learning with Time-Adaptive Privacy Spending

链接：https://iclr.cc/virtual/2025/poster/29388 abstract： Federated learning (FL) with differential privacy (DP) provides a framework for collaborative machine learning, enabling clients to train a shared model while adhering to strict privacy constraints. The framework allows each client to have an individual privacy guarantee, e.g., by adding different amounts of noise to each client's model updates. One underlying assumption is that all clients spend their privacy budgets uniformly over time (learning rounds). However, it has been shown in the literature that learning in early rounds typically focuses on more coarse-grained features that can be learned at lower signal-to-noise ratios while later rounds learn fine-grained features that benefit from higher signal-to-noise ratios. Building on this intuition, we propose a time-adaptive DP-FL framework that expends the privacy

budget non-uniformly across both time and clients. Our framework enables each client to save privacy budget in early rounds so as to be able to spend more in later rounds when additional accuracy is beneficial in learning more fine-grained features. We theoretically prove utility improvements in the case that clients with stricter privacy budgets spend budgets unevenly across rounds, compared to clients with more relaxed budgets, who have sufficient budgets to distribute their spend more evenly. Our practical experiments on standard benchmark datasets support our theoretical results and show that, in practice, our algorithms improve the privacy-utility trade-offs compared to baseline schemes.

1044. Solving Video Inverse Problems Using Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29542 abstract： Recently, diffusion model-based inverse problem solvers (DIS) have emerged as state-of-the-art approaches for addressing inverse problems, including image super-resolution, deblurring, inpainting, etc. However, their application to video inverse problems arising from spatio-temporal degradation remains largely unexplored due to the challenges in training video diffusion models.To address this issue, here we introduce an innovative video inverse solver that leverages only image diffusion models.Specifically, bydrawing inspiration from the success of the recent decomposed diffusion sampler (DDS), our method treats the time dimension of a video as the batch dimension of image diffusion models and solves spatio-temporal optimization problems within denoised spatio-temporal batches derived from each image diffusion model.Moreover, we introduce a batch-consistent diffusion sampling strategy that encourages consistency across batches by synchronizing the stochastic noise components in image diffusion models. Our approach synergistically combines batch-consistent sampling with simultaneous optimization of denoised spatio-temporal batches at each reverse diffusion step, resulting in a novel and efficient diffusion sampling strategy for video inverse problems.Experimental results demonstrate that our method effectively addresses various spatio-temporal degradations in video inverse problems, achieving state-of-the-art reconstructions.Project page: https://svi-diffusion.github.io/

1045. Breaking the Reclustering Barrier in Centroid-based Deep Clustering

链接：https://iclr.cc/virtual/2025/poster/28226 abstract： This work investigates an important phenomenon in centroid-based deep clustering (DC) algorithms: Performance quickly saturates after a period of rapid early gains. Practitioners commonly address early saturation with periodic reclustering, which we demonstrate to be insufficient to address performance plateaus. We call this phenomenon the “reclustering barrier” and empirically show when the reclustering barrier occurs, what its underlying mechanisms are, and how it is possible to Break the Reclustering Barrier with our algorithm BRB. BRB avoids early over-commitment to initial clusterings and enables continuous adaptation to reinitialized clustering targets while remaining conceptually simple. Applying our algorithm to widely-used centroid-based DC algorithms, we show that (1) BRB consistently improves performance across a wide range of clustering benchmarks, (2) BRB enables training from scratch, and (3) BRB performs competitively against state-of-the-art DC algorithms when combined with a contrastive loss. We release our code and pre-trained models at https://github.com/Probabilistic-and-Interactive-ML/breaking-the-reclustering-barrier .

1046. Should VLMs be Pre-trained with Image Data?

链接：https://iclr.cc/virtual/2025/poster/29734 abstract： Pre-trained LLMs that are further trained with image data perform well on vision-language tasks. While adding images during a second training phase effectively unlocks this capability, it is unclear how much of a gain or loss this two-step pipeline gives over VLMs which integrate images earlier into the training process. To investigate this, we train models spanning various datasets, scales, image-text ratios, and amount of pre-training done before introducing vision tokens.We then fine-tune these models and evaluate their downstream performance on a suite of vision-language and text-only tasks.We find that pre-training with a mixture of image and text data allows models to perform better on vision-language tasks while maintaining strong performance on text-only evaluations.On an average of 6 diverse tasks, we find that for a 1B model, introducing visual tokens 80\% of the way through pre-training results in a 2\% average improvement over introducing visual tokens to a fully pre-trained model.

1047. Efficient Alternating Minimization with Applications to Weighted Low Rank Approximation

链接：https://iclr.cc/virtual/2025/poster/28163 abstract： Weighted low rank approximation is a fundamental problem in numerical linear algebra, and it has many applications in machine learning. Given a matrix $M \in \mathbb{R}^{n \times n}$, a non-negative weight matrix $W \in \mathbb{R}_{\geq 0}^{n \times n}$, a parameter k, the goal is to output two matrices $X,Y\in \mathbb{R}^{n \times k}$ such that $\| W \circ (M - X Y^\top) \|_F$ is minimized, where $\circ$ denotes the Hadamard product. It naturally generalizes the well-studied low rank matrix completion problem. Such a problem is known to be NP-hard and even hard to approximate assuming the Exponential Time Hypothesis. Meanwhile, alternating minimization is a good heuristic solution for weighted low rank approximation. In particular, [Li, Liang and Risteski, ICML'16] shows that, under mild assumptions, alternating minimization does provide provable guarantees. In this work, we develop an efficient and robust framework for alternating minimization that allows the alternating updates to be computed approximately. For weighted low rank approximation, this improves the runtime of [Li, Liang and Risteski, ICML'16] from $\|W\|_0k^2$ to $\|W\|_0 k$ where $\|W\|_0$ denotes the number of nonzero entries of the weight matrix. At the heart of our framework is a high-accuracy multiple response regression solver together with a robust analysis of alternating minimization.

1048. On a Connection Between Imitation Learning and RLHF

链接：https://iclr.cc/virtual/2025/poster/31138 abstract： This work studies the alignment of large language models with preference data from an imitation learning perspective. We establish a close theoretical connection between reinforcement learning from human feedback RLHF and imitation learning (IL), revealing that RLHF implicitly performs imitation learning on the preference data distribution. Building on this connection, we propose DIL, a principled framework that directly optimizes the imitation learning objective. DIL provides a unified imitation learning perspective on alignment, encompassing existing alignment algorithms as special cases while naturally introducing new variants. By bridging IL and RLHF, DIL offers new insights into alignment with RLHF. Extensive experiments demonstrate that DIL outperforms existing methods on various challenging benchmarks.

1049. Competitive Fair Scheduling with Predictions

链接：https://iclr.cc/virtual/2025/poster/28648 abstract： Beyond the worst-case analysis of algorithms, the learning-augmented framework considers that an algorithm can leverage possibly imperfect predictions about the unknown variables to have guarantees tied to the prediction quality. We consider online non-clairvoyant scheduling to minimize the max-stretch under this framework, where the scheduler can access job size predictions. We present a family of algorithms: Relaxed-Greedy (RG) with an $O(\eta^3 \cdot \sqrt{P})$ competitive ratio, where η denotes the prediction error for job sizes and P the maximum job size ratio; Adaptive Relaxed-Greedy with an $O(\lambda^{0.5} \cdot \eta^{2.5} \cdot \sqrt{P})$ competitive ratio, where λ denotes the error for the minimum job size; Predictive Relaxed-Greedy with an $O(\lambda^{0.5} \cdot \varphi^{0.5} \cdot \eta \cdot \max \{ \eta, \varphi \} \cdot \sqrt{P})$ competitive ratio, where φ denotes the error for the maximum job size. We also present *${RG}^x$*, an algorithm that represents a trade-off between consistency and smoothness, with an $O(\eta^{2+2x} \cdot P^{1-x})$ competitive ratio. We introduce a general method using resource augmentation to bound robustness, resulting in *RR*-augmented *RG*, with a $(1 + \epsilon)$-speed $O(\min \{ \eta^3 \sqrt{P}, \frac{n}{\epsilon} \})$ competitive ratio. Finally, we conduct simulations on synthetic and real-world datasets to evaluate the practical performance of these algorithms.

1050. Bayesian Experimental Design Via Contrastive Diffusions

链接：https://iclr.cc/virtual/2025/poster/28775 abstract： Bayesian Optimal Experimental Design (BOED) is a powerful tool to reduce the cost of running a sequence of experiments.When based on the Expected Information Gain (EIG), design optimization corresponds to the maximization of some intractable expected contrast between prior and posterior distributions.Scaling this maximization to high dimensional and complex settings has been an issue due to BOED inherent computational complexity.In this work, we introduce an pooled posterior distribution with cost-effective sampling properties and provide a tractable access to the EIG contrast maximization via a new EIG gradient expression. Diffusion-based samplers are used to compute the dynamics of the pooled posterior and ideas from bi-level optimization are leveraged to derive an efficient joint sampling-optimization loop, without resorting to lower bound approximations of the EIG. The resulting efficiency gain allows to extend BOED to the well-tested generative capabilities of diffusion models. By incorporating generative models into the BOED framework, we expand its scope and its use in scenarios that were previously impractical. Numerical experiments and comparison with state-of-the-art methods show the potential of the approach.

1051. Estimation of single-cell and tissue perturbation effect in spatial transcriptomics via Spatial Causal Disentanglement

链接：https://iclr.cc/virtual/2025/poster/29512 abstract： Models of Virtual Cells and Virtual Tissues at single-cell resolution would allow us to test perturbations in silico and accelerate progress in tissue and cell engineering. However, most such models are not rooted in causal inference and as a result, could mistake correlation for causation.We introduce Celcomen, a novel generative graph neural network grounded in mathematical causality to disentangle intra- and inter-cellular gene regulation in spatial transcriptomics and single-cell data. Celcomen can also be prompted by perturbations to generate spatial counterfactuals, thus offering insights into experimentally inaccessible states, with potential applications in human health. We validate the model's disentanglement and identifiability through simulations, and demonstrate its counterfactual predictions in clinically relevant settings, including human glioblastoma and fetal spleen, recovering inflammation-related gene programs post immune system perturbation. Moreover, it supports mechanistic interpretability, as its parameters can be reverse-engineered from observed behavior, making it an accessible model for understanding both neural networks and complex biological systems.

1052. Learning a Neural Solver for Parametric PDEs to Enhance Physics-Informed Methods

链接：https://iclr.cc/virtual/2025/poster/28615 abstract： Physics-informed deep learning often faces optimization challenges due to the complexity of solving partial differential equations (PDEs), which involve exploring large solution spaces, require numerous iterations, and can lead to unstable training. These challenges arise particularly from the ill-conditioning of the optimization problem, caused by the differential terms in the loss function. To address these issues, we propose learning a solver, i.e., solving PDEs using a physics-informed iterative algorithm trained on data. Our method learns to condition a gradient descent algorithm that automatically adapts to each PDE instance, significantly accelerating and stabilizing the optimization process and enabling faster convergence of physics-aware models. Furthermore, while traditional physics-informed methods solve for a single PDE instance, our approach addresses parametric PDEs. Specifically, our method integrates the physical loss gradient with the PDE parameters to solve over a distribution of PDE parameters, including coefficients, initial conditions,

or boundary conditions. We demonstrate the effectiveness of our method through empirical experiments on multiple datasets, comparing training and test-time optimization performance.

1053. Second-Order Min-Max Optimization with Lazy Hessians

链接：https://iclr.cc/virtual/2025/poster/28680 abstract： This paper studies second-order methods for convex-concave minimax optimization. Monteiro & Svaiter (2012) proposed a method to solve the problem with an optimal iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ to find an ϵ-saddle point. However, it is unclear whether thecomputational complexity, $\mathcal{O}((N+ d^2) d \epsilon^{-2/3})$, can be improved. In the above, we follow Doikov et al. (2023) and assume the complexity of obtaining a first-order oracle as N and the complexity of obtaining a second-order oracle as dN. In this paper, we show that the computation cost can be reduced by reusing Hessian across iterations. Our methods take the overall computational complexity of $\tilde{\mathcal{O}}((N+d^2)(d+ d^{2/3}\epsilon^{-2/3}))$, which improves those of previous methods by a factor of $d^{1/3}$. Furthermore, we generalize our method to strongly-convex-strongly-concave minimax problems and establish the complexity of $\tilde{\mathcal{O}}((N+d^2) (d + d^{2/3} \kappa^{2/3}))$ when the condition number of the problem is κ, enjoying a similar speedup upon the state-of-the-art method. Numerical experiments on both real and synthetic datasets also verify the efficiency of our method.

1054. DenseMatcher: Learning 3D Semantic Correspondence for Category-Level Manipulation from a Single Demo

链接：https://iclr.cc/virtual/2025/poster/30743 abstract： Dense 3D correspondence can enhance robotic manipulation by enabling the generalization of spatial, functional, and dynamic information from one object to an unseen counterpart. Compared to shape correspondence, semantic correspondence is more effective in generalizing across different object categories. To this end, we present DenseMatcher, a method capable of computing 3D correspondences between in-the-wild objects that share similar structures. DenseMatcher first computes vertex features by projecting multiview 2D features onto meshes and refining them with a 3D network, and subsequently finds dense correspondences with the obtained features using functional map. In addition, we craft the first 3D matching dataset that contains colored object meshes across diverse categories. We demonstrate the downstream effectiveness of DenseMatcher in (i) robotic manipulation, where it achieves cross-instance and cross-category generalization on long-horizon complex manipulation tasks from observing only one demo; (ii) zero-shot color mapping between digital assets, where appearance can be transferred between different objects with relatable geometry. More details and demonstrations can be found at https://tea-lab.github.io/DenseMatcher/.

1055. 3DTrajMaster: Mastering 3D Trajectory for Multi-Entity Motion in Video Generation

链接：https://iclr.cc/virtual/2025/poster/30249 abstract： This paper aims to manipulate multi-entity 3D motions in video generation. Previous methods on controllable video generation primarily leverage 2D control signals to manipulate object motions and have achieved remarkable synthesis results. However, 2D control signals are inherently limited in expressing the 3D nature of object motions. To overcome this problem, we introduce 3DTrajMaster, a robust controller that regulates multi-entity dynamics in 3D space, given user-desired 6DoF pose (location and rotation) sequences of entities. At the core of our approach is a plug-and-play 3D-motion grounded object injector that fuses multiple input entities with their respective 3D trajectories through a gated self-attention mechanism. In addition, we exploit an injector architecture to preserve the video diffusion prior, which is crucial for generalization ability. To mitigate video quality degradation, we introduce a domain adaptor during training and employ an annealed sampling strategy during inference. To address the lack of suitable training data, we construct a 360-Motion Dataset, which first correlates collected 3D human and animal assets with GPT-generated trajectory and then captures their motion with 12 evenly-surround cameras on diverse 3D UE platforms. Extensive experiments show that 3DTrajMaster sets a new state-of-the-art in both accuracy and generalization for controlling multi-entity 3D motions. Project page: http://fuxiao0719.github.io/projects/3dtrajmaster

1056. Online-to-Offline RL for Agent Alignment

链接：https://iclr.cc/virtual/2025/poster/28165 abstract： Reinforcement learning (RL) has shown remarkable success in training agents to achieve high-performing policies, particularly in domains like Game AI where simulation environments enable efficient interactions. However, despite their success in maximizing these returns, such online-trained policies often fail to align with human preferences concerning actions, styles, and values. The challenge lies in efficiently adapting these online-trained policies to align with human preferences, given the scarcity and high cost of collecting human behavior data. In this work, we formalize the problem as online-to-offline RL and propose ALIGNment of Game AI to Preferences (ALIGN-GAP), an innovative approach for the alignment of well-trained game agents to human preferences. Our method features a carefully designed reward model that encodes human preferences from limited offline data and incorporates curriculum-based preference learning to align RL agents with targeted human preferences. Experiments across diverse environments and preference types demonstrate the performance of ALIGN-GAP, achieving effective alignment with human preferences.

1057. More RLHF, More Trust? On The Impact of Preference Alignment On Trustworthiness

链接：https://iclr.cc/virtual/2025/poster/30325 abstract： The trustworthiness of Large Language Models (LLMs) refers to the extent to which their outputs are reliable, safe, and ethically aligned, and it has become a crucial consideration alongside their cognitive performance. In practice, Reinforcement Learning From Human Feedback (RLHF) has been widely used to align LLMs with labeled human preferences, but its assumed effect on model trustworthiness hasn't been rigorously evaluated. To bridge this knowledge gap, this study investigates how models aligned with general-purpose preference data perform across five trustworthiness verticals: toxicity, stereotypical bias, machine ethics, truthfulness, and privacy. Our results demonstrate that RLHF on human preferences doesn't automatically guarantee trustworthiness, and reverse effects are often observed. Furthermore, we propose to adapt efficient influence function based data attribution methods to the RLHF setting to better understand the influence of fine-tuning data on individual trustworthiness benchmarks, and show its feasibility by providing our estimated attribution scores. Together, our results underscore the need for more nuanced approaches for model alignment from both the data and framework perspectives, and we hope this research will guide the community towards developing language models that are increasingly capable without sacrificing trustworthiness.

1058. URLOST: Unsupervised Representation Learning without Stationarity or Topology

链接：https://iclr.cc/virtual/2025/poster/29948 abstract： Unsupervised representation learning has seen tremendous progress. However, it is constrained by its reliance on domain specific stationarity and topology, a limitation not found in biological intelligence systems. For instance, unlike computer vision, human vision can process visual signals sampled from highly irregular and non-stationary sensors. We introduce a novel framework that learns from high-dimensional data without prior knowledge of stationarity and topology. Our model, abbreviated as URLOST, combines a learnable self-organizing layer, spectral clustering, and a masked autoencoder (MAE). We evaluate its effectiveness on three diverse data modalities including simulated biological vision data, neural recordings from the primary visual cortex, and gene expressions. Compared to state-of-the-art unsupervised learning methods like SimCLR and MAE, our model excels at learning meaningful representations across diverse modalities without knowing their stationarity or topology. It also outperforms other methods that are not dependent on these factors, setting a new benchmark in the field. We position this work as a step toward unsupervised learning methods capable of generalizing across diverse high-dimensional data modalities.

1059. Lambda-Skip Connections: the architectural component that prevents Rank Collapse

链接：https://iclr.cc/virtual/2025/poster/31171 abstract： Rank collapse, a phenomenon where embedding vectors in sequence modelsrapidly converge to a uniform token or equilibrium state, has recently gained at-tention in the deep learning literature. This phenomenon leads to reduced expres-sivity and potential training instabilities due to vanishing gradients. Empirical ev-idence suggests that architectural components like skip connections, LayerNorm,and MultiLayer Perceptrons (MLPs) play critical roles in mitigating rank collapse.While this issue is well-documented for transformers, alternative sequence mod-els, such as State Space Models (SSMs), which have recently gained prominence,have not been thoroughly examined for similar vulnerabilities. This paper extendsthe theory of rank collapse from transformers to SSMs using a unifying frame-work that captures both architectures. We introduce a modification in the skipconnection component, termed lambda-skip connections, that provides guaran-tees for rank collapse prevention. We present, via analytical results, a sufficientcondition to achieve the guarantee for all of the aforementioned architectures. Wealso study the necessity of this condition via ablation studies and analytical exam-ples. To our knowledge, this is the first study that provides a general guarantee toprevent rank collapse, and that investigates rank collapse in the context of SSMs,offering valuable understanding for both theoreticians and practitioners. Finally,we validate our findings with experiments demonstrating the crucial role of archi-tectural components in preventing rank collapse.

1060. Decision Tree Induction Through LLMs via Semantically-Aware Evolution

链接：https://iclr.cc/virtual/2025/poster/29435 abstract： Decision trees are a crucial class of models offering robust predictive performance and inherent interpretability across various domains, including healthcare, finance, and logistics. However, current tree induction methods often face limitations such as suboptimal solutions from greedy methods or prohibitive computational costs and limited applicability of exact optimization approaches.To address these challenges, we propose an evolutionary optimization method for decision tree induction based on genetic programming (GP). Our key innovation is the integration of semantic priors and domain-specific knowledge about the search space into the optimization algorithm.To this end, we introduce $\texttt{LLEGO}$, a framework that incorporates semantic priors into genetic search operators through the use of Large Language Models (LLMs), thereby enhancing search efficiency and targeting regions of the search space that yield decision trees with superior generalization performance. This is operationalized through novel genetic operators that work with structured natural language prompts, effectively utilizing LLMs as conditional generative models and sources of semantic knowledge. Specifically, we introduce $\textit{fitness-guided}$ crossover to exploit high-performing regions, and $\textit{diversity-guided}$ mutation for efficient global exploration of the search space. These operators are controlled by corresponding hyperparameters that enable a more nuanced balance between exploration and exploitation across the search space. Empirically, we demonstrate across various benchmarks that $\texttt{LLEGO}$ evolves superior-performing trees compared to existing tree induction methods, and exhibits significantly more efficient search performance compared to conventional GP approaches.

1061. Active Task Disambiguation with LLMs

链接：https://iclr.cc/virtual/2025/poster/30123 abstract： Despite the impressive performance of large language models (LLMs) across various benchmarks, their ability to address ambiguously specified problems—frequent in real-world interactions—remains underexplored. To address this gap, we introduce a formal definition of task ambiguity and frame the problem of task disambiguation through the lens of Bayesian Experimental Design. By posing clarifying questions, LLM agents can acquire additional task specifications, progressively narrowing the space of viable solutions and reducing the risk of generating unsatisfactory outputs. Yet, generating effective clarifying questions requires LLM agents to engage in a form of meta-cognitive reasoning, an ability LLMs may presently lack. Our proposed approach of active task disambiguation enables LLM agents to generate targeted questions maximizing the information gain. Effectively, this approach shifts the load from implicit to explicit reasoning about the space of viable solutions. Empirical results demonstrate that this form of question selection leads to more effective task disambiguation in comparison to approaches relying on reasoning solely within the space of questions.

1062. Examining Alignment of Large Language Models through Representative Heuristics: the case of political stereotypes

链接：https://iclr.cc/virtual/2025/poster/30828 abstract： Examining the alignment of large language models (LLMs) has become increasingly important, e.g., when LLMs fail to operate as intended. This study examines the alignment of LLMs with human values for the domain of politics. Prior research has shown that LLM-generated outputs can include political leanings and mimic the stances of political parties on various issues. However, the extent and conditions under which LLMs deviate from empirical positions are insufficiently examined. To address this gap, we analyze the factors that contribute to LLMs' deviations from empirical positions on political issues, aiming to quantify these deviations and identify the conditions that cause them. Drawing on findings from cognitive science about representativeness heuristics, i.e., situations where humans lean on representative attributes of a target group in a way that leads to exaggerated beliefs, we scrutinize LLM responses through this heuristics' lens. We conduct experiments to determine how LLMs inflate predictions about political parties, which results in stereotyping. We find that while LLMs can mimic certain political parties' positions, they often exaggerate these positions more than human survey respondents do. Also, LLMs tend to overemphasize representativeness more than humans. This study highlights the susceptibility of LLMs to representativeness heuristics, suggesting a potential vulnerability of LLMs that facilitates political stereotyping. We also test prompt-based mitigation strategies, finding that strategies that can mitigate representative heuristics in humans are also effective in reducing the influence of representativeness on LLM-generated responses.

1063. Knowledge Localization: Mission Not Accomplished? Enter Query Localization!

链接：https://iclr.cc/virtual/2025/poster/28033 abstract： Large language models (LLMs) store extensive factual knowledge, but the mechanisms behind how they store and express this knowledge remain unclear.The Knowledge Neuron (KN) thesis is a prominent theory for explaining these mechanisms. This theory is based on the Knowledge Localization (KL) assumption, which suggests that a fact can be localized to a few knowledge storage units, namely knowledge neurons. However, this assumption has two limitations: first, it may be too rigid regarding knowledge storage, and second, it neglects the role of the attention module in knowledge expression. In this paper, we first re-examine the KL assumption and demonstrate that its limitations do indeed exist. To address these, we then present two new findings, each targeting one of the limitations: one focusing on knowledge storage and the other on knowledge expression.We summarize these findings as Query Localization assumption and argue that the KL assumption can be viewed as a simplification of the QL assumption. Based on QL assumption, we further propose the Consistency-Aware KN modification method, which improves the performance of knowledge modification, further validating our new assumption. We conduct 39 sets of experiments, along with additional visualization experiments, to rigorously confirm our conclusions. Code will be made public soon.

1064. MIRAGE: Evaluating and Explaining Inductive Reasoning Process in Language Models

链接：https://iclr.cc/virtual/2025/poster/28046 abstract： Inductive reasoning is an essential capability for large language models (LLMs) to achieve higher intelligence, which requires the model to generalize rules from observed facts and then apply them to unseen examples. We present {\scshape Mirage}, a synthetic dataset that addresses the limitations of previous work, specifically the lack of comprehensive evaluation and flexible test data. In it, we evaluate LLMs' capabilities in both the inductive and deductive stages, allowing for flexible variation in input distribution, task scenario, and task difficulty to analyze the factors influencing LLMs' inductive reasoning. Based on these multi-faceted evaluations, we demonstrate that the LLM is a poor rule-based reasoner. In many cases, when conducting inductive reasoning, they do not rely on a correct rule to answer the unseen case. From the perspectives of different prompting methods, observation numbers, and task forms, models tend to consistently conduct correct deduction without correct inductive rules. Besides, we find that LLMs are good neighbor-based reasoners. In the inductive reasoning process, the model tends to focus on observed facts that are close to the current test example in feature space. By leveraging these similar examples, the model maintains strong inductive capabilities within a localized region, significantly improving its deductive performance.

1065. Controlled LLM Decoding via Discrete Auto-regressive Biasing

链接：https://iclr.cc/virtual/2025/poster/30436 abstract： Controlled text generation allows for enforcing user-defined constraints on large language model outputs, an increasingly important field as LLMs become more prevalent in everyday life. One common approach uses energy-based decoding, which defines a target distribution through an energy function that combines multiple constraints into a weighted average. However, these methods often struggle to balance fluency with constraint satisfaction, even with extensive tuning of the energy function's coefficients. In this paper, we identify that this suboptimal balance arises from sampling in continuous space rather than the natural discrete space of text tokens. To address this, we propose \emph{Discrete Auto-regressive Biasing}, a controlled decoding algorithm that leverages gradients while operating entirely in the discrete text domain.Specifically, we introduce a new formulation for controlled text generation by defining a joint distribution over the generated sequence and an auxiliary bias sequence. To efficiently sample from this joint distribution, we propose a Langevin-within-Gibbs sampling algorithm using gradient-based discrete MCMC. Our method significantly improves constraint satisfaction while maintaining comparable or better fluency, all with lower computational costs. We demonstrate the advantages of our controlled decoding method on sentiment control, language detoxification, and keyword-guided generation.

1066. ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time

链接：https://iclr.cc/virtual/2025/poster/29674 abstract： Vision Language Models (VLMs) have become essential backbones for multi-modal intelligence, yet significant safety challenges limit their real-world application. While textual inputs can often be effectively safeguarded, adversarial visual inputs can often easily bypass VLM defense mechanisms. Existing defense methods are either resource-intensive, requiring substantial data and compute, or fail to simultaneously ensure safety and usefulness in responses. To address these limitations, we propose a novel two-phase inference-time alignment framework, **E**valuating **T**hen **A**ligning (ETA): i) Evaluating input visual contents and output responses to establish a robust safety awareness in multimodal settings, and ii) Aligning unsafe behaviors at both shallow and deep levels by conditioning the VLMs' generative distribution with an interference prefix and performing sentence-level best-of-N to search the most harmless and helpful generation paths. Extensive experiments show that ETA outperforms baseline methods in terms of harmlessness, helpfulness, and efficiency, reducing the unsafe rate by 87.5\% in cross-modality attacks and achieving 96.6\% win-ties in GPT-4 helpfulness evaluation.

1067. Partially Observed Trajectory Inference using Optimal Transport and a Dynamics Prior

链接：https://iclr.cc/virtual/2025/poster/30243 abstract： Trajectory inference seeks to recover the temporal dynamics of a population from snapshots of its (uncoupled) temporal marginals, i.e. where observed particles are \emph{not} tracked over time. Prior works addressed this challenging problem under a stochastic differential equation (SDE) model with a gradient-driven drift in the observed space, introducing a minimum entropy estimator relative to the Wiener measure and a practical grid-free mean-field Langevin (MFL) algorithm using Schr\"odinger bridges. Motivated by the success of observable state space models in the traditional paired trajectory inference problem (e.g. target tracking), we extend the above framework to a class of latent SDEs in the form of \emph{observable state space models}. In this setting, we use partial observations to infer trajectories in the latent space under a specified dynamics model (e.g. the constant velocity/acceleration models from target tracking). We introduce the PO-MFL algorithm to solve this latent trajectory inference problem and provide theoretical guarantees to the partially observed setting. Experiments validate the robustness of our method and the exponential convergence of the MFL dynamics, and demonstrate significant outperformance over the latent-free baseline in key scenarios.

1068. Optimization by Parallel Quasi-Quantum Annealing with Gradient-Based Sampling

链接：https://iclr.cc/virtual/2025/poster/30713 abstract： Learning-based methods have gained attention as general-purpose solvers due to their ability to automatically learn problem-specific heuristics, reducing the need for manually crafted heuristics. However, these methods often face scalability challenges. To address these issues, the improved Sampling algorithm for Combinatorial Optimization (iSCO), using discrete Langevin dynamics, has been proposed, demonstrating better performance than several learning-based solvers. This study proposes a different approach that integrates gradient-based update through continuous relaxation, combined with Quasi-Quantum Annealing (QQA). QQA smoothly transitions the objective function, starting from a simple convex function, minimized at half-integral values, to the original objective function, where the relaxed variables are minimized only in the discrete space. Furthermore, we incorporate parallel run communication leveraging GPUs to enhance exploration capabilities and accelerate convergence. Numerical experiments demonstrate that our method is a competitive general-purpose solver, achieving performance comparable to iSCO and learning-based solvers across various benchmark problems. Notably, our method exhibits superior speed-quality trade-offs for large-scale instances compared to iSCO, learning-based solvers, commercial solvers, and specialized algorithms.

1069. OSCAR: Operating System Control via State-Aware Reasoning and Re-Planning

链接：https://iclr.cc/virtual/2025/poster/29394 abstract： Large language models (LLMs) and large multimodal models (LMMs) have shown great potential in automating complex tasks like web browsing and gaming. However, their ability to generalize across diverse applications remains limited, hindering broader utility. To address this challenge, we present OSCAR: Operating

System Control via state-Aware reasoning and Re-planning. OSCAR is a generalist agent designed to autonomously navigate and interact with various desktop and mobile applications through standardized controls, such as mouse and keyboard inputs, while processing screen images to fulfill user commands. OSCAR translates human instructions into executable Python code, enabling precise control over graphical user interfaces (GUIs). To enhance stability and adaptability, OSCAR operates as a state machine, equipped with error-handling mechanisms and dynamic task re-planning, allowing it to efficiently adjust to real-time feedback and exceptions. We demonstrate OSCAR's effectiveness through extensive experiments on diverse benchmarks across desktop and mobile platforms, where it transforms complex workflows into simple natural language commands, significantly boosting user productivity. Our code will be open-source upon publication.

1070. Exploring the Effectiveness of Object-Centric Representations in Visual Question Answering: Comparative Insights with Foundation Models

链接：https://iclr.cc/virtual/2025/poster/30471 abstract： Object-centric (OC) representations, which model visual scenes as compositions of discrete objects, have the potential to be used in various downstream tasks to achieve systematic compositional generalization and facilitate reasoning. However, these claims have yet to be thoroughly validated empirically.Recently, foundation models have demonstrated unparalleled capabilities across diverse domains, from language to computer vision, positioning them as a potential cornerstone of future research for a wide range of computational tasks.In this paper, we conduct an extensive empirical study on representation learning for downstream Visual Question Answering (VQA), which requires an accurate compositional understanding of the scene. We thoroughly investigate the benefits and trade-offs of OC models and alternative approaches including large pre-trained foundation models on both synthetic and real-world data, ultimately identifying a promising path to leverage the strengths of both paradigms. The extensiveness of our study, encompassing over 600 downstream VQA models and 15 different types of upstream representations, also provides several additional insights that we believe will be of interest to the community at large.

1071. Bayesian Optimization of Antibodies Informed by a Generative Model of Evolving Sequences

链接：https://iclr.cc/virtual/2025/poster/30425 abstract： To build effective therapeutics, biologists iteratively mutate antibody sequences to improve binding and stability. Proposed mutations can be informed by previous measurements or by learning from large antibody databases to predict only typical antibodies. Unfortunately, the space of typical antibodies is enormous to search, and experiments often fail to find suitable antibodies on a budget. We introduce Clone-informed Bayesian Optimization (CloneBO), a Bayesian optimization procedure that efficiently optimizes antibodies in the lab by teaching a generative model how our immune system optimizes antibodies. Our immune system makes antibodies by iteratively evolving specific portions of their sequences to bind their target strongly and stably, resulting in a set of related, evolving sequences known as a clonal family. We train a large language model, CloneLM, on hundreds of thousands of clonal families and use it to design sequences with mutations that are most likely to optimize an antibody within the human immune system. We propose to guide our designs to fit previous measurements with a twisted sequential Monte Carlo procedure. We show that CloneBO optimizes antibodies substantially more efficiently than previous methods in realistic in silico experiments and designs stronger and more stable binders in in vitro wet lab experiments.

1072. Diverse Policies Recovering via Pointwise Mutual Information Weighted Imitation Learning

链接：https://iclr.cc/virtual/2025/poster/30905 abstract： Recovering a spectrum of diverse policies from a set of expert trajectories is an important research topic in imitation learning. After determining a latent style for a trajectory, previous diverse polices recovering methods usually employ a vanilla behavioral cloning learning objective conditioned on the latent style, treating each state-action pair in the trajectory with equal importance. Based on an observation that in many scenarios, behavioral styles are often highly relevant with only a subset of state-action pairs, this paper presents a new principled method in diverse polices recovering. In particular, after inferring or assigning a latent style for a trajectory, we enhance the vanilla behavioral cloning by incorporating a weighting mechanism based on pointwise mutual information.This additional weighting reflects the significance of each state-action pair's contribution to learning the style, thus allowing our method to focus on state-action pairs most representative of that style.We provide theoretical justifications for our new objective, and extensive empirical evaluations confirm the effectiveness of our method in recovering diverse polices from expert data.

1073. Advancing Graph Generation through Beta Diffusion

链接：https://iclr.cc/virtual/2025/poster/27807 abstract： Diffusion models have excelled in generating natural images and are now being adapted to a variety of data types, including graphs. However, conventional models often rely on Gaussian or categorical diffusion processes, which can struggle to accommodate the mixed discrete and continuous components characteristic of graph data. Graphs typically feature discrete structures and continuous node attributes that often exhibit rich statistical patterns, including sparsity, bounded ranges, skewed distributions, and long-tailed behavior. To address these challenges, we introduce Graph Beta Diffusion (GBD), a generative model specifically designed to handle the diverse nature of graph data. GBD leverages a beta diffusion process, effectively modeling both continuous and discrete elements. Additionally, we propose a modulation technique that enhances the realism of generated graphs by stabilizing critical graph topology while

maintaining flexibility for other components. GBD competes strongly with existing models across multiple general and biochemical graph benchmarks, showcasing its ability to capture the intricate balance between discrete and continuous features inherent in real-world graph data. Our PyTorch code is available at https://github.com/xinyangATK/GraphBetaDiffusion.

1074. GenSE: Generative Speech Enhancement via Language Models using Hierarchical Modeling

链接：https://iclr.cc/virtual/2025/poster/31185 abstract： Semantic information refers to the meaning conveyed through words, phrases, and contextual relationships within a given linguistic structure. Humans can leverage semantic information, such as familiar linguistic patterns and contextual cues, to reconstruct incomplete or masked speech signals in noisy environments. However, existing speech enhancement (SE) approaches often overlook the rich semantic information embedded in speech, which is crucial for improving intelligibility, speaker consistency, and overall quality of enhanced speech signals. To enrich the SE model with semantic information, we employ language models as an efficient semantic learner and propose a comprehensive framework tailored for language model-based speech enhancement, called GenSE. Specifically, we approach SE as a conditional language modeling task rather than a continuous signal regression problem defined in existing works. This is achieved by tokenizing speech signals into semantic tokens using a pre-trained self-supervised model and into acoustic tokens using a custom-designed single-quantizer neural codec model. To improve the stability of language model predictions, we propose a hierarchical modeling method that decouples the generation of clean semantic tokens and clean acoustic tokens into two distinct stages. Moreover, we introduce a token chain prompting mechanism during the acoustic token generation stage to ensure timbre consistency throughout the speech enhancement process. Experimental results on benchmark datasets demonstrate that our proposed approach outperforms state-of-the-art SE systems in terms of speech quality and generalization capability. Codes and demos are publicly available at https://anonymous.4open.science/w/gen-se-7F52/.

1075. Certifying Counterfactual Bias in LLMs

链接：https://iclr.cc/virtual/2025/poster/30226 abstract： Large Language Models (LLMs) can produce biased responses that can cause representational harms. However, conventional studies are insufficient to thoroughlyevaluate biases across LLM responses for different demographic groups (a.k.a.counterfactual bias), as they do not scale to large number of inputs and do notprovide guarantees. Therefore, we propose the first framework, LLMCert-B thatcertifies LLMs for counterfactual bias on distributions of prompts. A certificateconsists of high-confidence bounds on the probability of unbiased LLM responsesfor any set of counterfactual prompts - prompts differing by demographic groups,sampled from a distribution. We illustrate counterfactual bias certification fordistributions of counterfactual prompts created by applying prefixes sampled fromprefix distributions, to a given set of prompts. We consider prefix distributions consisting random token sequences, mixtures of manual jailbreaks, and perturbationsof jailbreaks in LLM’s embedding space. We generate non-trivial certificates forSOTA LLMs, exposing their vulnerabilities over distributions of prompts generatedfrom computationally inexpensive prefix distributions.

1076. A Black Swan Hypothesis: The Role of Human Irrationality in AI Safety

链接：https://iclr.cc/virtual/2025/poster/30807 abstract： Black swan events are statistically rare occurrences that carry extremely high risks. A typical view of defining black swan events is heavily assumed to originate from an unpredictable time-varying environments; however, the community lacks a comprehensive definition of black swan events. To this end, this paper challenges that the standard view is incomplete and claims that high-risk, statistically rare events can also occur in unchanging environments due to human misperception of their value and likelihood, which we call as spatial black swan event. We first carefully categorize black swan events, focusing on spatial black swan events, and mathematically formalize the definition of black swan events. We hope these definitions can pave the way for the development of algorithms to prevent such events by rationally correcting human perception.

1077. Enhancing Cognition and Explainability of Multimodal Foundation Models with Self-Synthesized Data

链接：https://iclr.cc/virtual/2025/poster/28527 abstract： Large Multimodal Models (LMMs), or Vision-Language Models (VLMs), have shown impressive capabilities in a wide range of visual tasks. However, they often struggle with fine-grained visual reasoning, failing to identify domain-specific objectives and provide justifiable explanations for their predictions. To address the above challenge, we propose a novel visual rejection sampling framework to improve the cognition and explainability of LMMs using self-synthesized data. Specifically, visual fine-tuning requires images, queries, and target answers. Our approach begins by synthesizing interpretable answers that include human-verifiable visual features. These features are based on expert-defined concepts, and carefully selected based on their alignment with the image content. After each round of fine-tuning, we apply a reward model-free filtering mechanism to select the highest-quality interpretable answers for the next round of tuning. This iterative process of synthetic data generation and fine-tuning progressively improves the model's ability to generate accurate and reasonable explanations. Experimental results demonstrate the effectiveness of our method in improving both the accuracy and explainability of specialized visual classification tasks.

1078. AIMS.au: A Dataset for the Analysis of Modern Slavery Countermeasures in Corporate Statements

链接：https://iclr.cc/virtual/2025/poster/27711 abstract： Despite over a decade of legislative efforts to address modern slavery in the supply chains of large corporations, the effectiveness of government oversight remains hampered by the challenge of scrutinizing thousands of statements annually. While Large Language Models (LLMs) can be considered a well established solution for the automatic analysis and summarization of documents, recognizing concrete modern slavery countermeasures taken by companies and differentiating those from vague claims remains a challenging task. To help evaluate and fine-tune LLMs for the assessment of corporate statements, we introduce a dataset composed of 5,731 modern slavery statements taken from the Australian Modern Slavery Register and annotated at the sentence level. This paper details the construction steps for the dataset that include the careful design of annotation specifications, the selection and preprocessing of statements, and the creation of high-quality annotation subsets for effective model evaluations. To demonstrate our dataset's utility, we propose a machine learning methodology for the detection of sentences relevant to mandatory reporting requirements set by the Australian Modern Slavery Act. We then follow this methodology to benchmark modern language models under zero-shot and supervised learning settings.

1079. MQuAKE-Remastered: Multi-Hop Knowledge Editing Can Only Be Advanced with Reliable Evaluations

链接：https://iclr.cc/virtual/2025/poster/28478 abstract： Large language models (LLMs) can give out erroneous answers to factually rooted questions either as a result of undesired training outcomes or simply because the world has moved on after a certain knowledge cutoff date. Under such scenarios, knowledge editing often comes to the rescue by delivering efficient patches for such erroneous answers without significantly altering the rest, where many editing methods have seen reasonable success when the editing targets are simple and direct (e.g., ``what club does Lionel Messi currently play for?'').However, knowledge fragments like this are often deeply intertwined in the real world, making effectively propagating the editing effect to non-directly related questions a practical challenge (to entertain an extreme example: "What car did the wife of the owner of the club that Messi currently plays for used to get to school in the 80s?"). Prior arts have coined this task as multi-hop knowledge editing with the most popular dataset being MQuAKE, serving as the sole evaluation benchmark for many later proposed editing methods due to the expensive nature of constructing knowledge editing datasets at scale. In this work, we reveal that up to 33\% or 76\% of \mquake{}'s questions and ground truth labels are, in fact, corrupted in various fashions due to some unintentional clerical or procedural oversights. Our work provides a detailed audit of MQuAKE's error pattern and a comprehensive fix without sacrificing its dataset capacity. Additionally, we benchmarked almost all proposed MQuAKE-evaluated editing methods on our post-fix dataset, MQuAKE-Remastered. We observe that many methods try to overfit the original MQuAKE by exploiting some dataset idiosyncrasies of MQuAKE. We provide a guideline on how to approach such datasets faithfully and show that a simple, minimally invasive approach — GWalk — can offer beyond SOTA editing performance without such exploitation. The MQuAKE-Remastered datasets and utilities are available at huggingface.co/datasets/henryzhongsc/MQuAKE-Remastered and github.com/henryzhongsc/MQuAKE-Remastered, respectively.

1080. Laplace Sample Information: Data Informativeness Through a Bayesian Lens

链接：https://iclr.cc/virtual/2025/poster/28251 abstract： Accurately estimating the informativeness of individual samples in a dataset is an important objective in deep learning, as it can guide sample selection, which can improve model efficiency and accuracy by removing redundant or potentially harmful samples. We propose $\text{\textit{Laplace Sample Information}}$ (LSI) measure of sample informativeness grounded in information theory widely applicable across model architectures and learning settings.LSI leverages a Bayesian approximation to the weight posterior and the KL divergence to measure the change in the parameter distribution induced by a sample of interest from the dataset.We experimentally show that LSI is effective in ordering the data with respect to typicality, detecting mislabeled samples, measuring class-wise informativeness, and assessing dataset difficulty.We demonstrate these capabilities of LSI on image and text data in supervised and unsupervised settings.Moreover, we show that LSI can be computed efficiently through probes and transfers well to the training of large models.

1081. S4M: S4 for multivariate time series forecasting with Missing values

链接：https://iclr.cc/virtual/2025/poster/30554 abstract： Multivariate time series data play a pivotal role in a wide range of real-world applications, such as finance, healthcare, and meteorology, where accurate forecasting is critical for informed decision-making and proactive interventions. However, the presence of block missing data introduces significant challenges, often compromising the performance of predictive models. Traditional two-step approaches, which first impute missing values and then perform forecasting, are prone to error accumulation, particularly in complex multivariate settings characterized by high missing ratios and intricate dependency structures. In this work, we introduce S4M, an end-to-end time series forecasting framework that seamlessly integrates missing data handling into the Structured State Space Sequence (S4) model architecture. Unlike conventional methods that treat imputation as a separate preprocessing step, S4M leverages the latent space of S4 models to directly recognize and represent missing data patterns, thereby more effectively capturing the underlying temporal and multivariate dependencies. Our framework comprises two key components: the Adaptive Temporal Prototype Mapper (ATPM) and the Missing-Aware Dual Stream S4 (MDS-S4). The ATPM employs a prototype bank to derive robust and informative representations from historical data patterns, while the MDS-S4 processes these representations alongside missingness masks as dual input streams to enable accurate forecasting. Through extensive empirical evaluations on diverse real-world datasets, we demonstrate that S4M consistently achieves state-of-the-art performance. These results underscore the efficacy of our integrated approach in handling missing data, showcasing its robustness and superiority over traditional imputation-based

methods. Our findings highlight the potential of S4M to advance reliable time series forecasting in practical applications, offering a promising direction for future research and deployment. Code is available at https://github.com/WINTERWEEL/S4M.git.

1082. Weak to Strong Generalization for Large Language Models with Multi-capabilities

链接：https://iclr.cc/virtual/2025/poster/29903 abstract： As large language models (LLMs) grow in sophistication, some of their capabilities surpass human abilities, making it essential to ensure their alignment with human values and intentions, i.e., Superalignment. This superalignment challenge is particularly critical for complex tasks, as annotations provided by humans, as weak supervisors, may be overly simplistic, incomplete, or incorrect. Previous work has demonstrated the potential of training a strong model using the weak dataset generated by a weak model as weak supervision. However, these studies have been limited to a single capability. In this work, we conduct extensive experiments to investigate weak to strong generalization for LLMs with multi-capabilities. The experiments reveal that different capabilities tend to remain relatively independent in this generalization, and the effectiveness of weak supervision is significantly impacted by the quality and diversity of the weak datasets. Moreover, the self-bootstrapping of the strong model leads to performance degradation due to its overconfidence and the limited diversity of its generated dataset. To address these issues, we proposed a novel training framework using reward models to select valuable data, thereby providing weak supervision for strong model training. In addition, we propose a two-stage training method on both weak and selected datasets to train the strong model. Experimental results demonstrate our method significantly improves the weak to strong generalization with multi-capabilities.

1083. On the Benefits of Memory for Modeling Time-Dependent PDEs

链接：https://iclr.cc/virtual/2025/poster/28374 abstract： Data-driven techniques have emerged as a promising alternative to traditional numerical methods for solving PDEs. For time-dependent PDEs, many approaches are Markovian---the evolution of the trained system only depends on the current state, and not the past states. In this work, we investigate the benefits of using memory for modeling time-dependent PDEs: that is, when past states are explicitly used to predict the future. Motivated by the Mori-Zwanzig theory of model reduction, we theoretically exhibit examples of simple (even linear) PDEs, in which a solution that uses memory is arbitrarily better than a Markovian solution. Additionally, we introduce Memory Neural Operator (MemNO), a neural operator architecture that combines recent state space models (specifically, S4) and Fourier Neural Operators (FNOs) to effectively model memory. We empirically demonstrate that when the PDEs are supplied in low resolution or contain observation noise at train and test time, MemNO significantly outperforms the baselines without memory---with up to $6 \times$ reduction in test error. Furthermore, we show that this benefit is particularly pronounced when the PDE solutions have significant high-frequency Fourier modes (e.g., low-viscosity fluid dynamics) and we construct a challenging benchmark dataset consisting of such PDEs.

1084. Studying the Interplay Between the Actor and Critic Representations in Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28067 abstract： Extracting relevant information from a stream of high-dimensional observations is a central challenge for deep reinforcement learning agents. Actor-critic algorithms add further complexity to this challenge, as it is often unclear whether the same information will be relevant to both the actor and the critic. To this end, we here explore the principles that underlie effective representations for the actor and for the critic in on-policy algorithms. We focus our study on understanding whether the actor and critic will benefit from separate, rather than shared, representations. Our primary finding is that when separated, the representations for the actor and critic systematically specialise in extracting different types of information from the environment---the actor's representation tends to focus on action-relevant information, while the critic's representation specialises in encoding value and dynamics information. We conduct a rigourous empirical study to understand how different representation learning approaches affect the actor and critic's specialisations and their downstream performance, in terms of sample efficiency and generation capabilities. Finally, we discover that a separated critic plays an important role in exploration and data collection during training. Our code, trained models and data are accessible at https://github.com/francelico/deac-rep.

1085. The Case for Cleaner Biosignals: High-fidelity Neural Compressor Enables Transfer from Cleaner iEEG to Noisier EEG

链接：https://iclr.cc/virtual/2025/poster/29124 abstract： All data modalities are not created equal, even when the signal they measure comes from the same source. In the case of the brain, two of the most important data modalities are the scalp electroencephalogram (EEG), and the intracranial electroencephalogram (iEEG). iEEG benefits from a higher signal-to-noise ratio (SNR), as it measures the electrical activity directly in the brain, while EEG is noisier and has lower spatial and temporal resolutions. Nonetheless, both EEG and iEEG are important sources of data for human neurology, from healthcare to brain–machine interfaces. They are used by human experts, supported by deep learning (DL) models, to accomplish a variety of tasks, such as seizure detection and motor imagery classification. Although the differences between EEG and iEEG are well understood by human experts, the performance of DL models across these two modalities remains under-explored. To help characterize the importance of clean data on the performance of DL models, we propose BrainCodec, a high-fidelity EEG and iEEG neural compressor. We find that training BrainCodec on iEEG and then transferring to EEG yields higher reconstruction

quality than training on EEG directly. In addition, we also find that training BrainCodec on both EEG and iEEG improves fidelity when reconstructing EEG. Our work indicates that data sources with higher SNR, such as iEEG, provide better performance across the board also in the medical time-series domain. This finding is consistent with reports coming from natural language processing, where clean data sources appear to have an outsized effect on the performance of the DL model overall. BrainCodec also achieves up to a 64x compression on iEEG and EEG without a notable decrease in quality. BrainCodec markedly surpasses current state-of-the-art compression models both in final compression ratio and in reconstruction fidelity. We also evaluate the fidelity of the compressed signals objectively on a seizure detection and a motor imagery task performed by standard DL models. Here, we find that BrainCodec achieves a reconstruction fidelity high enough to ensure no performance degradation on the downstream tasks. Finally, we collect the subjective assessment of an expert neurologist, that confirms the high reconstruction quality of BrainCodec in a realistic scenario. The code is available at https://github.com/IBM/eeg-ieeg-brain-compressor.

1086. Monitoring Latent World States in Language Models with Propositional Probes

链接：https://iclr.cc/virtual/2025/poster/31228 abstract： Language models (LMs) are susceptible to bias, sycophancy, backdoors, and other tendencies that lead to unfaithful responses to the input context. Interpreting internal states of LMs could help monitor and correct unfaithful behavior. We hypothesize that LMs faithfully represent their input contexts in a latent world model, and we seek to extract these latent world states as logical propositions. For example, given the input context ``Greg is a nurse. Laura is a physicist.'', we aim to decode the propositions WorksAs(Greg, nurse) and WorksAs(Laura, physicist) from the model's internal activations. To do so we introduce *propositional probes*, which compositionally extract lexical concepts from token activations and bind them into propositions. Key to this is identifying a *binding subspace* in which bound tokens have high similarity (Greg $\leftrightarrow$ nurse) but unbound ones do not (Greg $\not\leftrightarrow$ physicist). Despite only being trained on linguistically simple English templates, we find that propositional probes generalize to inputs written as short stories and translated to Spanish. Moreover, in three settings where LMs respond unfaithfully to the input context---prompt injections, backdoor attacks, and gender bias--- the decoded propositions remain faithful. This suggests that LMs often encode a faithful world model but decode it unfaithfully, which motivates the search for better interpretability tools for monitoring LMs.

1087. BAMDP Shaping: a Unified Framework for Intrinsic Motivation and Reward Shaping

链接：https://iclr.cc/virtual/2025/poster/28031 abstract： Intrinsic motivation and reward shaping guide reinforcement learning (RL) agents by adding pseudo-rewards, which can lead to useful emergent behaviors. However, they can also encourage counterproductive exploits, e.g., fixation with noisy TV screens. Here we provide a theoretical model which anticipates these behaviors, and provides broad criteria under which adverse effects can be bounded. We characterize all pseudo-rewards as reward shaping in Bayes-Adaptive Markov Decision Processes (BAMDPs), which formulates the problem of learning in MDPs as an MDP over the agent's knowledge. Optimal exploration maximizes BAMDP state value, which we decompose into the value of the information gathered and the prior value of the physical state. Psuedo-rewards guide RL agents by rewarding behavior that increases these value components, while they hinder exploration when they align poorly with the actual value. We extend potential-based shaping theory to prove BAMDP Potential-based shaping Functions (BAMPFs) are immune to reward-hacking (convergence to behaviors maximizing composite rewards to the detriment of real rewards) in meta-RL, and show empirically how a BAMPF helps a meta-RL agent learn optimal RL algorithms for a Bernoulli Bandit domain. We finally prove that BAMPFs with bounded monotone increasing potentials also resist reward-hacking in the regular RL setting. We show that it is straightforward to retrofit or design new pseudo-reward terms in this form, and provide an empirical demonstration in the Mountain Car environment.

1088. Do LLMs have Consistent Values?

链接：https://iclr.cc/virtual/2025/poster/30729 abstract： Large Language Models (LLM) technology is rapidly advancing towards human- like dialogue. Values are fundamental drivers of human behavior, yet research on the values expressed in LLM-generated text remains limited. While prior work has begun to explore value ranking in LLMs, the crucial aspect of value correlation – the interrelationship and consistency between different values – has been largely un-examined. Drawing on established psychological theories of human value structure, this paper investigates whether LLMs exhibit human-like value correlations within a single session, reflecting a coherent “persona”. Our findings reveal that standard prompting methods fail to produce human-consistent value correlations. However, we demonstrate that a novel prompting strategy (referred to as "Value Anchoring"), significantly improves the alignment of LLM value correlations with human data. Furthermore, we analyze the mechanism by which Value Anchoring achieves this effect. These results not only deepen our understanding of value representation in LLMs but also introduce new methodologies for evaluating consistency and human-likeness in LLM responses, highlighting the importance of explicit value prompting for generating human-aligned outputs.

1089. Multilevel Generative Samplers for Investigating Critical Phenomena

链接：https://iclr.cc/virtual/2025/poster/29254 abstract： Investigating critical phenomena or phase transitions is of high interest in physics and chemistry, for which Monte Carlo (MC) simulations, a crucial tool for numerically analyzing macroscopic properties of given systems, are often hindered by an emerging divergence of correlation length---known as scale invariance at

criticality (SIC) in the renormalization group theory. SIC causes the system to behave the same at any length scale, from which many existing sampling methods suffer: long-range correlations cause critical slowing down in Markov chain Monte Carlo (MCMC), and require intractably large receptive fields for generative samplers. In this paper, we propose a Renormalization-informed Generative Critical Sampler (RiGCS)---a novel sampler specialized for near-critical systems, where SIC is leveraged as an advantage rather than a nuisance. Specifically, RiGCS builds on MultiLevel Monte Carlo (MLMC) with Heat Bath (HB) algorithms, which perform ancestral sampling from low-resolution to high-resolution lattice configurations with site wise-independent conditional HB sampling. Although MLMC-HB is highly efficient under exact SIC, it suffers from a low acceptance rate under slight SIC violation. Notably, SIC violation always occurs in finite-size systems, and may induce long-range and higher-order interactions in the renormalized distributions, which are not considered by independent HB samplers. RiGCS enhances MLMC-HB by replacing a part of the conditional HB sampler with generative models that capture those residual interactions and improve the sampling efficiency. Our experiments show that the effective sample size of RiGCS is a few orders of magnitude higher than state-of-the-art generative model baselines in sampling configurations for 128×128 two-dimensional Ising systems. SIC also allows us to adopt a specialized sequential training protocol with model transfer, which significantly accelerates training.

1090. Sensor-Invariant Tactile Representation

链接：https://iclr.cc/virtual/2025/poster/29640 abstract： High-resolution tactile sensors have become critical for embodied perception and robotic manipulation. However, a key challenge in the field is the lack of transferability between sensors due to design and manufacturing variations, which result in significant differences in tactile signals. This limitation hinders the ability to transfer models or knowledge learned from one sensor to another. To address this, we introduce a novel method for extracting Sensor-Invariant Tactile Representations (SITR), enabling zero-shot transfer across optical tactile sensors. Our approach utilizes a transformer-based architecture trained on a diverse dataset of simulated sensor designs, allowing it to generalize to new sensors in the real world with minimal calibration. Experimental results demonstrate the method’s effectiveness across various tactile sensing applications, facilitating data and model transferability for future advancements in the field.

1091. Gaussian Ensemble Belief Propagation for Efficient Inference in High-Dimensional, Black-box Systems

链接：https://iclr.cc/virtual/2025/poster/29758 abstract： Efficient inference in high-dimensional models is a central challenge in machine learning.We introduce the Gaussian Ensemble Belief Propagation (GEnBP) algorithm, which combines the strengths of the Ensemble Kalman Filter (EnKF) and Gaussian Belief Propagation (GaBP) to address this challenge.GEnBP updates ensembles of prior samples into posterior samples by passing low-rank local messages over the edges of a graphical model, enabling efficient handling of high-dimensional states, parameters, and complex, noisy, black-box generative processes.By utilizing local message passing within a graphical model structure, GEnBP effectively manages complex dependency structures and remains computationally efficient even when the ensemble size is much smaller than the inference dimension --- a common scenario in spatiotemporal modeling, image processing, and physical model inversion.We demonstrate that GEnBP can be applied to various problem structures, including data assimilation, system identification, and hierarchical models, and show through experiments that it outperforms existing belief propagation methods in terms of accuracy and computational efficiency.Supporting code is available at https://github.com/danmackinlay/GEnBP}{github.com/danmackinlay/GEnBP

1092. Rethinking Invariance Regularization in Adversarial Training to Improve Robustness-Accuracy Trade-off

链接：https://iclr.cc/virtual/2025/poster/29949 abstract： Adversarial training often suffers from a robustness-accuracy trade-off, where achieving high robustness comes at the cost of accuracy.One approach to mitigate this trade-off is leveraging invariance regularization, which encourages model invariance under adversarial perturbations; however, it still leads to accuracy loss.In this work, we closely analyze the challenges of using invariance regularization in adversarial training and understand how to address them.Our analysis identifies two key issues: (1) a "gradient conflict" between invariance and classification objectives, leading to suboptimal convergence, and (2) the mixture distribution problem arising from diverged distributions between clean and adversarial inputs.To address these issues, we propose Asymmetric Representation-regularized Adversarial Training (ARAT), which incorporates asymmetric invariance loss with stop-gradient operation and a predictor to avoid gradient conflict, and a split-BatchNorm (BN) structure to resolve the mixture distribution problem.Our detailed analysis demonstrates that each component effectively addresses the identified issues, offering novel insights into adversarial defense.ARAT shows superiority over existing methods across various settings. Finally, we discuss the implications of our findings to knowledge distillation-based defenses, providing a new perspective on their relative successes.

1093. 3D-SPATIAL MULTIMODAL MEMORY

链接：https://iclr.cc/virtual/2025/poster/29300 abstract： We present 3D Spatial MultiModal Memory (M3), a multimodal memory system designed to retain information about medium-sized static scenes through video sources for visual perception. By integrating 3D Gaussian Splatting techniques with foundation models, M3 builds a multimodal memory capable of rendering feature representations across granularities, encompassing a wide range of knowledge. In our exploration, we identify two key challenges in previous works on feature splatting: (1) computational constraints in storing high-dimensional features for each Gaussian primitive, and (2) misalignment or information loss between distilled features and foundation model features. To

address these challenges, we propose M3 with key components of principal scene components and Gaussian memory attention, enabling efficient training and inference. To validate M3, we conduct comprehensive quantitative evaluations of feature similarity and downstream tasks, as well as qualitative visualizations to highlight the pixel trace of Gaussian memory attention. Our approach encompasses a diverse range of foundation models, including vision-language models (VLMs), perception models, and large multimodal and language models (LMMs/LLMs). Furthermore, to demonstrate real-world applicability, we deploy M3's feature field in indoor scenes on a quadruped robot. Notably, we claim that M3 is the first work to address the core compression challenges in 3D feature distillation.

1094. Adaptive Transformer Programs: Bridging the Gap Between Performance and Interpretability in Transformers

链接：https://iclr.cc/virtual/2025/poster/29383 abstract： Balancing high performance with interpretability in increasingly powerful Transformer-based models remains a challenge. While mechanistic interpretability aims to specify neural network computations in explicit, pseudocode-like formats, existing methods often involve laborious manual analysis or struggle to fully elucidate learned internal algorithms. Recent efforts to build intrinsically interpretable models have introduced considerable expressivity and optimization challenges. This work introduces Adaptive Transformer Programs, an enhanced framework building upon RASP language and Transformer Programs to create more robust and interpretable models. The proposed method increases expressivity by redesigning two primary attention modules to improve categorical and numerical reasoning capabilities. To overcome optimization hurdles, we introduce a novel reparameterization scheme that enhances the exploration-exploitation trade-off during training. We validate our approach through extensive experiments on diverse tasks, including in-context learning, algorithmic problems (e.g., sorting and Dyck languages), and NLP benchmarks such as named entity recognition and text classification. Results demonstrate that Adaptive Transformer Programs substantially narrow the performance gap between black-box Transformers and interpretable models, enhancing transparency. This work advances the development of high-performing, transparent AI systems for critical applications, addressing crucial ethical concerns in AI development.

1095. Controllable Blur Data Augmentation Using 3D-Aware Motion Estimation

链接：https://iclr.cc/virtual/2025/poster/29336 abstract： Existing realistic blur datasets provide insufficient variety in scenes and blur patterns to be trained, while expanding data diversity demands considerable time and effort due to complex dual-camera systems. To address the challenge, data augmentation can be an effective way to artificially increase data diversity. However, existing methods on this line are typically designed to estimate motions from a 2D perspective, e.g., estimating 2D non-uniform kernels disregarding 3D aspects of blur modeling, which leads to unrealistic motion patterns due to the fact that camera and object motions inherently arise in 3D space. In this paper, we propose a 3D-aware blur synthesizer capable of generating diverse and realistic blur images for blur data augmentation. Specifically, we estimate 3D camera positions within the motion blur interval, generate the corresponding scene images, and aggregate them to synthesize a realistic blur image. Since the 3D camera positions projected onto the 2D image plane inherently lie in 2D space, we can represent the 3D transformation as a combination of 2D transformation and projected 3D residual component. This allows for 3D transformation without requiring explicit depth measurements, as the 3D residual component is directly estimated via a neural network. Furthermore, our blur synthesizer allows for controllable blur data augmentation by modifying blur magnitude, direction, and scenes, resulting in diverse blur images. As a result, our method significantly improves deblurring performance, making it more practical for real-world scenarios.

1096. A Causal Lens for Learning Long-term Fair Policies

链接：https://iclr.cc/virtual/2025/poster/28199 abstract： Fairness-aware learning studies the development of algorithms that avoid discriminatory decision outcomes despite biased training data. While most studies have concentrated on immediate bias in static contexts, this paper highlights the importance of investigating long-term fairness in dynamic decision-making systems while simultaneously considering instantaneous fairness requirements. In the context of reinforcement learning, we propose a general framework where long-term fairness is measured by the difference in the average expected qualification gain that individuals from different groups could obtain. Then, through a causal lens, we decompose this metric into three components that represent the direct impact, the delayed impact, as well as the spurious effect the policy has on the qualification gain. We analyze the intrinsic connection between these components and an emerging fairness notion called benefit fairness that aims to control the equity of outcomes in decision-making. Finally, we develop a simple yet effective approach for balancing various fairness notions.

1097. Mechanism and Emergence of Stacked Attention Heads in Multi-Layer Transformers

链接：https://iclr.cc/virtual/2025/poster/28194 abstract： In this paper, I introduce the retrieval problem, a simple yet common reasoning task that can be solved only by transformers with a minimum number of layers, which grows logarithmically with the input size. I empirically show that large language models can solve the task under different prompting formulations without any fine-tuning. To understand how transformers solve the retrieval problem, I train several transformers on a minimal formulation.

Successful learning occurs only under the presence of an implicit curriculum. I uncover the learned mechanisms by studying the attention maps in the trained transformers. I also study the training process, uncovering that attention heads always emerge in a specific sequence guided by the implicit curriculum.

1098. Optimizing 4D Gaussians for Dynamic Scene Video from Single Landscape Images

链接：https://iclr.cc/virtual/2025/poster/30162 abstract： To achieve realistic immersion in landscape images, fluids such as water and clouds need to move within the image while revealing new scenes from various camera perspectives. Recently, a field called dynamic scene video has emerged, which combines single image animation with 3D photography. These methods use pseudo 3D space, implicitly represented with Layered Depth Images (LDIs). LDIs separate a single image into depth-based layers, which enables elements like water and clouds to move within the image while revealing new scenes from different camera perspectives. However, as landscapes typically consist of continuous elements, including fluids, the representation of a 3D space separates a landscape image into discrete layers, and it can lead to diminished depth perception and potential distortions depending on camera movement. Furthermore, due to its implicit modeling of 3D space, the output may be limited to videos in the 2D domain, potentially reducing their versatility. In this paper, we propose representing a complete 3D space for dynamic scene video by modeling explicit representations, specifically 4D Gaussians, from a single image. The framework is focused on optimizing 3D Gaussians by generating multi-view images from a single image and creating 3D motion to optimize 4D Gaussians. The most important part of proposed framework is consistent 3D motion estimation, which estimates common motion among multi-view images to bring the motion in 3D space closer to actual motions. As far as we know, this is the first attempt that considers animation while representing a complete 3D space from a single landscape image. Our model demonstrates the ability to provide realistic immersion in various landscape images through diverse experiments and metrics. Extensive experimental results are https://cvsp-lab.github.io/ICLR2025_3D-MOM/.

1099. ComaDICE: Offline Cooperative Multi-Agent Reinforcement Learning with Stationary Distribution Shift Regularization

链接：https://iclr.cc/virtual/2025/poster/30931 abstract： Offline reinforcement learning (RL) has garnered significant attention for its ability to learn effective policies from pre-collected datasets without the need for further environmental interactions. While promising results have been demonstrated in single-agent settings, offline multi-agent reinforcement learning (MARL) presents additional challenges due to the large joint state-action space and the complexity of multi-agent behaviors. A key issue in offline RL is the distributional shift, which arises when the target policy being optimized deviates from the behavior policy that generated the data. This problem is exacerbated in MARL due to the interdependence between agents' local policies and the expansive joint state-action space. Prior approaches have primarily addressed this challenge by incorporating regularization in the space of either Q-functions or policies. In this work, we propose a novel type of regularizer in the space of stationary distributions to address the distributional shift more effectively. Our algorithm, ComaDICE, provides a principled framework for offline cooperative MARL to correct the stationary distribution of the global policy, which is then leveraged to derive local policies for individual agents. Through extensive experiments on the offline multi-agent MuJoCo and StarCraft II benchmarks, we demonstrate that ComaDICE achieves superior performance compared to state-of-the-art offline MARL methods across nearly all tasks.

1100. Global Identifiability of Overcomplete Dictionary Learning via L1 and Volume Minimization

链接：https://iclr.cc/virtual/2025/poster/30993 abstract： We propose a novel formulation for dictionary learning with an overcomplete dictionary, i.e., when the number of atoms is larger than the dimension of the dictionary. The proposed formulation consists of a weighted sum of ℓ_1 norms of the rows of the sparse coefficient matrix plus the log of the matrix volume of the dictionary matrix. The main contribution of this work is to show that this novel formulation guarantees global identifiability of the overcomplete dictionary, under a mild condition that the sparse coefficient matrix satisfies a strong scattering condition in the hypercube. Furthermore, if every column of the coefficient matrix is sparse and the dictionary guarantees ℓ_1 recovery, then the coefficient matrix is identifiable as well. This is a major breakthrough for not only dictionary learning but also general matrix factorization models as identifiability is guaranteed even when the latent dimension is higher than the ambient dimension. We also provide a probabilistic analysis and show that if the sparse coefficient matrix is generated from the widely adopted sparse-Gaussian model, then the $m\times k$ overcomplete dictionary is globally identifiable if the sample size is bigger than a constant times $(k^2/m)\log(k^2/m)$ with overwhelming probability. Finally, we propose an algorithm based on alternating minimization to solve the new proposed formulation.

1101. ControlAR: Controllable Image Generation with Autoregressive Models

链接：https://iclr.cc/virtual/2025/poster/30567 abstract： Autoregressive (AR) models have reformulated image generation as next-token prediction, demonstrating remarkable potential and emerging as strong competitors to diffusion models. However, control-to-image generation, akin to ControlNet, remains largely unexplored within AR models. Although a natural approach, inspired by advancements in Large Language Models, is to tokenize control images into tokens and prefill them into the autoregressive model before decoding image tokens, it still falls short in generation quality compared to ControlNet and suffers

from inefficiency. To this end, we introduce ControlAR, an efficient and effective framework for integrating spatial controls into autoregressive image generation models. Firstly, we explore control encoding for AR models and propose a lightweight control encoder to transform spatial inputs (e.g., canny edges or depth maps) into control tokens. Then ControlAR exploits the conditional decoding method to generate the next image token conditioned on the per-token fusion between control and image tokens, similar to positional encodings. Compared to prefilling tokens, using conditional decoding significantly strengthens the control capability of AR models but also maintains the model efficiency. Furthermore, the proposed ControlAR surprisingly empowers AR models with arbitrary-resolution image generation via conditional decoding and specific controls. Extensive experiments can demonstrate the controllability of the proposed ControlAR for the autoregressive control-to-image generation across diverse inputs, including edges, depths, and segmentation masks. Furthermore, both quantitative and qualitative results indicate that ControlAR surpasses previous state-of-the-artcontrollable diffusion models, e.g., ControlNet++.

1102. Fat-to-Thin Policy Optimization: Offline Reinforcement Learning with Sparse Policies

链接：https://iclr.cc/virtual/2025/poster/29606 abstract： Sparse continuous policies are distributions that can choose some actions at random yet keep strictly zero probability for the other actions, which are radically different from the Gaussian.They have important real-world implications, e.g. in modeling safety-critical tasks like medicine.The combination of offline reinforcement learning and sparse policies provides a novel paradigm that enables learning completely from logged datasets a safety-aware sparse policy. However, sparse policies can cause difficulty with the existing offline algorithms which require evaluating actions that fall outside of the current support.In this paper, we propose the first offline policy optimization algorithm that tackles this challenge: Fat-to-Thin Policy Optimization (FtTPO).Specifically, we maintain a fat (heavy-tailed) proposal policy that effectively learns from the dataset and injects knowledge to a thin (sparse) policy, which is responsible for interacting with the environment.We instantiate FtTPO with the general q-Gaussian family that encompasses both heavy-tailed and sparse policies and verify that it performs favorably in a safety-critical treatment simulation and the standard MuJoCo suite.Our code is available at https://github.com/lingweizhu/fat2thin.

1103. MaskBit: Embedding-free Image Generation via Bit Tokens

链接：https://iclr.cc/virtual/2025/poster/31452 abstract： Masked transformer models for class-conditional image generation have become a compelling alternative to diffusion models. Typically comprising two stages - an initial VQGAN model for transitioning between latent space and image space, and a subsequent Transformer model for image generation within latent space - these frameworks offer promising avenues for image synthesis. In this study, we present two primary contributions: Firstly, an empirical and systematic examination of VQGANs, leading to a modernized VQGAN. Secondly, a novel embedding-free generation network operating directly on bit tokens - a binary quantized representation of tokens with rich semantics. The first contribution furnishes a transparent, reproducible, and high-performing VQGAN model, enhancing accessibility and matching the performance of current state-of-the-art methods while revealing previously undisclosed details. The second contribution demonstrates that embedding-free image generation using bit tokens achieves a new state-of-the-art FID of 1.52 on the ImageNet 256×256 benchmark, with a compact generator model of mere 305M parameters. The code for this project is available on https://github.com/markweberdev/maskbit.

1104. Connectome Mapping: Shape-Memory Network via Interpretation of Contextual Semantic Information

链接：https://iclr.cc/virtual/2025/poster/29747 abstract： Contextual semantic information plays a pivotal role in the brain's visual interpretation of the surrounding environment. When processing visual information, electrical signals within synapses facilitate the dynamic activation and deactivation of synaptic connections, guided by the contextual semantic information associated with different objects. In the realm of Artificial Intelligence (AI), neural networks have emerged as powerful tools to emulate complex signaling systems, enabling tasks such as classification and segmentation by understanding visual information. However, conventional neural networks have limitations in simulating the conditional activation and deactivation of synapses, collectively known as the connectome, a comprehensive map of neural connections in the brain. Additionally, the pixel-wise inference mechanism of conventional neural networks failed to account for the explicit utilization of contextual semantic information in the prediction process. To overcome these limitations, we developed a novel neural network, dubbed the Shape Memory Network (SMN), which excels in two key areas: (1) faithfully emulating the intricate mechanism of the brain's connectome, and (2) explicitly incorporating contextual semantic information during the inference process. The SMN memorizes the structure suitable for contextual semantic information and leverages this structure at the inference phase. The structural transformation emulates the conditional activation and deactivation of synaptic connections within the connectome. Rigorous experimentation carried out across a range of semantic segmentation benchmarks demonstrated the outstanding performance of the SMN, highlighting its superiority and effectiveness. Furthermore, our pioneering network on connectome emulation reveals the immense potential of the SMN for next-generation neural networks.

1105. Mind the Gap: Examining the Self-Improvement Capabilities of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28441 abstract： Self-improvement is a mechanism in Large Language Model (LLM)

pre-training, post-training and test-time inference. We explore a framework where the model verifies its own outputs, filters or reweights data based on this verification, and distills the filtered data. Despite several empirical successes, a fundamental understanding is still lacking. In this work, we initiate a comprehensive, modular and controlled study on LLM self-improvement. We provide a mathematical formulation for self-improvement, which is largely governed by a quantity which we formalize as the generation-verification gap. Through experiments with various model families and tasks, we discover a scaling phenomenon of self-improvement -- a variant of the generation-verification gap scales monotonically with the model pre-training flops. We also examine when self-improvement is possible, an iterative self-improvement procedure, and ways to improve its performance. Our findings not only advance understanding of LLM self-improvement with practical implications, but also open numerous avenues for future research into its capabilities and boundaries.

1106. Towards Understanding the Robustness of Diffusion-Based Purification: A Stochastic Perspective

链接：https://iclr.cc/virtual/2025/poster/28104 abstract： Diffusion-Based Purification (DBP) has emerged as an effective defense mechanism against adversarial attacks. The success of DBP is often attributed to the forward diffusion process, which reduces the distribution gap between clean and adversarial images by adding Gaussian noise. Although this explanation is theoretically grounded, the precise contribution of this process to robustness remains unclear. In this paper, through a systematic investigation, we propose that the intrinsic stochasticity in the DBP procedure is the primary factor driving robustness. To explore this hypothesis, we introduce a novel Deterministic White-Box (DW-box) evaluation protocol to assess robustness in the absence of stochasticity, and analyze attack trajectories and loss landscapes. Our results suggest that DBP models primarily leverage stochasticity to evade effective attack directions, and that their ability to purify adversarial perturbations can be weak. To further enhance the robustness of DBP models, we propose Adversarial Denoising Diffusion Training (ADDT), which incorporates classifier-guided adversarial perturbations into diffusion training, thereby strengthening the models' ability to purify adversarial perturbations. Additionally, we propose Rank-Based Gaussian Mapping (RBGM) to improve the compatibility of perturbations with diffusion models. Experimental results validate the effectiveness of ADDT. In conclusion, our study suggests that future research on DBP can benefit from the perspective of decoupling stochasticity-based and purification-based robustness.

1107. When LLMs Play the Telephone Game: Cultural Attractors as Conceptual Tools to Evaluate LLMs in Multi-turn Settings

链接：https://iclr.cc/virtual/2025/poster/28880 abstract： As large language models (LLMs) start interacting with each other and generating an increasing amount of text online, it becomes crucial to better understand how information is transformed as it passes from one LLM to the next. While significant research has examined individual LLM behaviors, existing studies have largely overlooked the collective behaviors and information distortions arising from iterated LLM interactions. Small biases, negligible at the single output level, risk being amplified in iterated interactions, potentially leading the content to evolve towards attractor states. In a series of telephone game experiments, we apply a transmission chain design borrowed from the human cultural evolution literature: LLM agents iteratively receive, produce, and transmit texts from the previous to the next agent in the chain. By tracking the evolution of text toxicity, positivity, difficulty, and length across transmission chains, we uncover the existence of biases and attractors, and study their dependence on the initial text, the instructions, language model, and model size. For instance, we find that more open-ended instructions lead to stronger attraction effects compared to more constrained tasks. We also find that different text properties display different sensitivity to attraction effects, with toxicity leading to stronger attractors than length. These findings highlight the importance of accounting for multi-step transmission dynamics and represent a first step towards a more comprehensive understanding of LLM cultural dynamics.

1108. Probabilistic Learning to Defer: Handling Missing Expert Annotations and Controlling Workload Distribution

链接：https://iclr.cc/virtual/2025/poster/27652 abstract： Recent progress in machine learning research is gradually shifting its focus towards human-AI cooperation due to the advantages of exploiting the reliability of human experts and the efficiency of AI models. One of the promising approaches in human-AI cooperation is learning to defer (L2D), where the system analyses the input data and decides to make its own decision or defer to human experts. Although L2D has demonstrated state-of-the-art performance, in its standard setting, L2D entails a severe limitation: all human experts must annotate the whole training dataset of interest, resulting in a time-consuming and expensive annotation process that can subsequently influence the size and diversity of the training set. Moreover, the current L2D does not have a principled way to control workload distribution among human experts and the AI classifier, which is critical to optimise resource allocation. We, therefore, propose a new probabilistic modelling approach inspired by the mixture-of-experts, where the Expectation - Maximisation algorithm is leverage to address the issue of missing expert's annotations. Furthermore, we introduce a constraint, which can be solved efficiently during the E-step, to control the workload distribution among human experts and the AI classifier. Empirical evaluation on synthetic and real-world datasets shows that our proposed probabilistic approach performs competitively, or surpasses previously proposed methods assessed on the same benchmarks.

1109. CPSample: Classifier Protected Sampling for Guarding Training Data During Diffusion

链接：https://iclr.cc/virtual/2025/poster/30007 abstract： Diffusion models have a tendency to exactly replicate their training data, especially when trained on small datasets. Most prior work has sought to mitigate this problem by imposing differential privacy constraints or masking parts of the training data, resulting in a notable substantial decrease in image quality. We present CPSample, a method that modifies the sampling process to prevent training data replication while preserving image quality. CPSample utilizes a classifier that is trained to overfit on random binary labels attached to the training data. CPSample then uses classifier guidance to steer the generation process away from the set of points that can be classified with high certainty, a set that includes the training data. CPSample achieves FID scores of 4.97 and 2.97 on CIFAR-10 and CelebA-64, respectively, without producing exact replicates of the training data. Unlike prior methods intended to guard the training images, CPSample only requires training a classifier rather than retraining a diffusion model, which is computationally cheaper. Moreover, our technique provides diffusion models with greater robustness against membership inference attacks, wherein an adversary attempts to discern which images were in the model's training dataset. We show that CPSample behaves like a built-in rejection sampler, and we demonstrate its capabilities to prevent mode collapse in Stable Diffusion.

1110. STAR: Synthesis of Tailored Architectures

链接：https://iclr.cc/virtual/2025/poster/30205 abstract： Iterative improvement of model architectures is fundamental to deep learning: Transformers first enabled scaling, and recent advances in model hybridization have pushed the quality-efficiency frontier. However, optimizing architectures remains challenging and expensive, with a variety of automated or manual approaches that fall short, due to limited progress in the design of search spaces and due to the simplicity of resulting patterns and heuristics. In this work, we propose a new approach for the synthesis of tailored architectures (STAR). Our approach combines a novel search space based on the theory of linear input-varying systems, supporting a hierarchical numerical encoding into architecture genomes. STAR genomes are automatically refined and recombined with gradient-free, evolutionary algorithms to optimize for multiple model quality and efficiency metrics. Using STAR, we optimize large populations of new architectures, leveraging diverse computational units and interconnection patterns, improving over highly-optimized Transformers and striped hybrid models on the frontier of quality, parameter size, and inference cache for autoregressive language modeling.

1111. RetroInText: A Multimodal Large Language Model Enhanced Framework for Retrosynthetic Planning via In-Context Representation Learning

链接：https://iclr.cc/virtual/2025/poster/30129 abstract： Development of robust and effective strategies for retrosynthetic planning requires a deep understanding of the synthesis process. A critical step in achieving this goal is accurately identifying synthetic intermediates. Current machine learning-based methods often overlook the valuable context from the overall route, focusing only on predicting reactants from the product, requiring cost annotations for every reaction step, and ignoring the multi-faced nature of molecular, resulting in inaccurate synthetic route predictions. Therefore, we introduce RetroInText, an advanced end-to-end framework based on a multimodal Large Language Model (LLM), featuring in-context learning with TEXT descriptions of synthetic routes. First, RetroInText including ChatGPT presents detailed descriptions of the reaction procedure. It learns the distinct compound representations in parallel with corresponding molecule encoders to extract multi-modal representations including 3D features. Subsequently, we propose an attention-based mechanism that offers a fusion module to complement these multi-modal representations with in-context learning and a fine-tuned language model for a single-step model. As a result, RetroInText accurately represents and effectively captures the complex relationship between molecules and the synthetic route. In experiments on the USPTO pathways dataset RetroBench, RetroInText outperforms state-of-the-art methods, achieving up to a 5% improvement in Top-1 test accuracy, particularly for long synthetic routes. These results demonstrate the superiority of RetroInText by integrating with context information over routes. They also demonstrate its potential for advancing pathway design and facilitating the development of organic chemistry. Code is available at https://github.com/guofei-tju/RetroInText.

1112. PhyloLM: Inferring the Phylogeny of Large Language Models and Predicting their Performances in Benchmarks

链接：https://iclr.cc/virtual/2025/poster/28195 abstract： This paper introduces PhyloLM, a method adapting phylogenetic algorithms to Large Language Models (LLMs) to explore whether and how they relate to each other and to predict their performance characteristics. Our method calculates a phylogenetic distance metric based on the similarity of LLMs' output. The resulting metric is then used to construct dendrograms, which satisfactorily capture known relationships across a set of 111 open-source and 45 closed models. Furthermore, our phylogenetic distance predicts performance in standard benchmarks, thus demonstrating its functional validity and paving the way for a time and cost-effective estimation of LLM capabilities. To sum up, by translating population genetic concepts to machine learning, we propose and validate a tool to evaluate LLM development, relationships and capabilities, even in the absence of transparent training information.

1113. Logically Consistent Language Models via Neuro-Symbolic Integration

链接：https://iclr.cc/virtual/2025/poster/30826 abstract： Current large language models (LLMs) are far from reliable: they are prone to generate non-factual information and, more crucially, to contradict themselves when prompted to reason about relations

between real entities of the world. These problems are currently addressed with large scale fine-tuning or by delegating consistent reasoning to external tools. In this work, we strive for a middle ground and leverage a training objective based on a principled neuro-symbolic loss that teaches a LLM to be consistent with external knowledge in the form of a set of facts and rules. Fine-tuning with such a loss on a limited set of facts enables our LLMs to be more logically consistent than previous baselines for a given constraint. Our approach also allows to easily combine multiple logical constraints at once in a principled way, delivering LLMs that are more consistent w.r.t. all the selected rules. Moreover, our method allows LLMs to extrapolate to unseen but semantically similar factual knowledge, represented in unseen datasets, more systematically.

1114. Improving Text-to-Image Consistency via Automatic Prompt Optimization

链接：https://iclr.cc/virtual/2025/poster/31459 abstract： Impressive advances in text-to-image (T2I) generative models have yielded a plethora of high performing models which are able to generate aesthetically appealing, photorealistic images. Despite the progress, these models still struggle to produce images that are consistent with the input prompt, oftentimes failing to capture object quantities, relations and attributes properly. Existing solutions to improve prompt-image consistency suffer from the following challenges: (1) they oftentimes require model fine-tuning, (2) they only focus on nearby prompt samples, and (3) they are affected by unfavorable trade-offs among image quality, representation diversity, and prompt-image consistency. In this paper, we address these challenges and introduce a T2I optimization-by-prompting framework, OPT2I, which leverages a large language model (LLM) to improve prompt-image consistency in T2I models. Our framework starts from a user prompt and iteratively generates revised prompts with the goal of maximizing a consistency score. Our extensive validation on two datasets, MSCOCO and PartiPrompts, shows that OPT2I can boost the initial consistency score by up to 24.9% in terms of DSG score while preserving the FID and increasing the recall between generated and real data. Our work paves the way toward building more reliable and robust T2I systems by harnessing the power of LLMs.

1115. OvercookedV2: Rethinking Overcooked for Zero-Shot Coordination

链接：https://iclr.cc/virtual/2025/poster/28745 abstract： AI agents hold the potential to transform everyday life by helping humans achieve their goals.To do this successfully, agents need to be able to coordinate with novel partners without prior interaction, a setting known as zero-shot coordination (ZSC).Overcooked has become one of the most popular benchmarks for evaluating coordination capabilities of AI agents and learning algorithms.In this work, we investigate the origins of ZSC challenges in Overcooked.We introduce a state augmentation mechanism which mixes states that might be encountered when paired with unknown partners into the training distribution, reducing the out-of-distribution challenge associated with ZSC.We show that independently trained agents under this algorithm coordinate successfully in Overcooked.Our results suggest that ZSC failure can largely be attributed to poor state coverage under self-play rather than more sophisticated coordination challenges. The Overcooked environment is therefore not suitable as a ZSC benchmark.To address these shortcomings, we introduce OvercookedV2, a new version of the benchmark, which includes asymmetric information and stochasticity, facilitating the creation of interesting ZSC scenarios.To validate OvercookedV2, we conduct experiments demonstrating that mere exhaustive state coverage is insufficient to coordinate well. Finally, we use OvercookedV2 to build a new range of coordination challenges, including ones that require test time protocol formation, and we demonstrate the need for new coordination algorithms that can adapt online.We hope that OvercookedV2 will help benchmark the next generation of ZSC algorithms and advance collaboration between AI agents and humans.

1116. SFESS: Score Function Estimators for k-Subset Sampling

链接：https://iclr.cc/virtual/2025/poster/28261 abstract： Are score function estimators a viable approach to learning with k-subset sampling? Sampling k-subsets is a fundamental operation that is not amenable to differentiable parametrization, impeding gradient-based optimization. Previous work has favored approximate pathwise gradients or relaxed sampling, dismissing score function estimators because of their high variance. Inspired by the success of score function estimators in variational inference and reinforcement learning, we revisit them for k-subset sampling. We demonstrate how to efficiently compute the distribution's score function using a discrete Fourier transform and reduce the estimator's variance with control variates. The resulting estimator provides both k-hot samples and unbiased gradient estimates while being applicable to non-differentiable downstream models, unlike existing methods. We validate our approach experimentally and find that it produces results comparable to those of recent state-of-the-art pathwise gradient estimators across a range of tasks.

1117. ET-SEED: EFFICIENT TRAJECTORY-LEVEL SE(3) EQUIVARIANT DIFFUSION POLICY

链接：https://iclr.cc/virtual/2025/poster/29807 abstract： Imitation learning, e.g., diffusion policy, has been proven effective in various robotic manipulation tasks.However, extensive demonstrations are required for policy robustness and generalization.To reduce the demonstration reliance, we leverage spatial symmetry and propose ET-SEED, an efficient trajectory-level SE(3) equivariant diffusion model for generating action sequences in complex robot manipulation tasks.Further, previous equivariant diffusion models require the per-step equivariance in the Markov process, making it difficult to learn policy under such strong constraints.We theoretically extend equivariant Markov kernels and simplify the condition of equivariant diffusion process, thereby significantly improving training efficiency for trajectory-level SE(3) equivariant diffusion policy in an end-to-end manner.We evaluate ET-SEED on representative robotic manipulation tasks, involving rigid body, articulated and deformable

object.Experiments demonstrate superior data efficiency and manipulation proficiency of our proposed method,as well as its ability to generalize to unseen configurations with only a few demonstrations. Website: https://et-seed.github.io/

1118. Generating Less Certain Adversarial Examples Improves Robust Generalization

链接：https://iclr.cc/virtual/2025/poster/31455 abstract： This paper revisits the robust overfitting phenomenon of adversarial training. Observing that models with better robust generalization performance are less certain in predicting adversarially generated training inputs, we argue that overconfidence in predicting adversarial examples is a potential cause. Therefore, we propose a formal definition of adversarial certainty that captures the variance of the model's predicted logits on adversarial examples and hypothesize that generating adversarial examples after the optimization of decreasing adversarial certainty improves robust generalization. Our theoretical analysis of synthetic distributions characterizes the connection between adversarial certainty and robust generalization. Accordingly, built upon the notion of adversarial certainty, we develop a general method to search for models that can generate training-time adversarial inputs with reduced certainty, while maintaining the model's capability in distinguishing adversarial examples. Extensive experiments on image benchmarks demonstrate that our method effectively learns models with consistently improved robustness and mitigates robust overfitting, confirming the importance of generating less certain adversarial examples for robust generalization. Our implementations are available as open-source code at: \url{https://github.com/TrustMLRG/AdvCertainty}.

1119. Discretization-invariance? On the Discretization Mismatch Errors in Neural Operators

链接：https://iclr.cc/virtual/2025/poster/30126 abstract： In recent years, neural operators have emerged as a prominent approach for learning mappings between function spaces, such as the solution operators of parametric PDEs. A notable example is the Fourier Neural Operator (FNO), which models the integral kernel as a convolution operator and uses the Convolution Theorem to learn the kernel directly in the frequency domain. The parameters are decoupled from the resolution of the data, allowing the FNO to take inputs of different resolutions.However, training at a lower resolution and inferring at a finer resolution does not guarantee consistent performance, nor can fine details, present only in fine-scale data, be learned solely from coarse data. In this work, we address this misconception by defining and examining the discretization mismatch error: the discrepancy between the outputs of the neural operator when using different discretizations of the input data. We demonstrate that neural operators may suffer from discretization mismatch errors that hinder their effectiveness when inferred on data with resolutions different from that of the training data or when trained on data with varying resolutions. As neural operators underpin many critical cross-resolution scientific tasks, such as climate modeling and fluid dynamics, understanding discretization mismatch errors is essential. Based on our findings, we propose a Cross-Resolution Operator-learning Pipeline that is free of aliasing and discretization mismatch errors, enabling efficient cross-resolution and multi-spatial-scale learning, and resulting in superior performance.

1120. Select before Act: Spatially Decoupled Action Repetition for Continuous Control

链接：https://iclr.cc/virtual/2025/poster/29767 abstract： Reinforcement Learning (RL) has achieved remarkable success in various continuous control tasks, such as robot manipulation and locomotion.Different to mainstream RL which makes decisions at individual steps, recent studies have incorporated action repetition into RL, achieving enhanced action persistence with improved sample efficiency and superior performance.However, existing methods treat all action dimensions as a whole during repetition, ignoring variations among them.This constraint leads to inflexibility in decisions, which reduces policy agility with inferior effectiveness. In this work, we propose a novel repetition framework called SDAR, which implements Spatially Decoupled Action Repetition through performing closed-loop act-or-repeat selection for each action dimension individually.SDAR achieves more flexible repetition strategies, leading to an improved balance between action persistence and diversity.Compared to existing repetition frameworks, SDAR is more sample efficient with higher policy performance and reduced action fluctuation.Experiments are conducted on various continuous control scenarios, demonstrating the effectiveness of spatially decoupled repetition design proposed in this work.

1121. Score-based free-form architectures for high-dimensional Fokker-Planck equations

链接：https://iclr.cc/virtual/2025/poster/30928 abstract： Deep learning methods incorporate PDE residuals as the loss function for solving Fokker-Planck equations, and usually impose the proper normalization condition to avoid a trivial solution. However, soft constraints require careful balancing of multi-objective loss functions, and specific network architectures may limit representation capacity under hard constraints. In this paper, we propose a novel framework: Fokker-Planck neural network (FPNN) that adopts a score PDE loss to decouple the score learning and the density normalization into two stages. Our method allows free-form network architectures to model the unnormalized density and strictly satisfy normalization constraints by post-processing. We demonstrate the effectiveness on various high-dimensional steady-state Fokker-Planck (SFP) equations, achieving superior accuracy and over a 20$\times$ speedup compared to state-of-the-art methods. Without any labeled data, FPNNs achieve the mean absolute percentage error (MAPE) of 11.36%, 13.87% and 12.72% for 4D Ring, 6D Unimodal and

6D Multi-modal problems respectively, requiring only 256, 980, and 980 parameters. Experimental results highlights the potential as a universal fast solver for handling more than 20-dimensional SFP equations, with great gains in efficiency, accuracy, memory and computational resource usage.

1122. ACTIVE: Offline Reinforcement Learning via Adaptive Imitation and In-sample V-Ensemble

链接：https://iclr.cc/virtual/2025/poster/28239 abstract： Offline reinforcement learning (RL) aims to learn from static datasets and thus faces the challenge of value estimation errors for out-of-distribution actions. The in-sample learning scheme addresses this issue by performing implicit TD backups that does not query the values of unseen actions. However, pre-existing in-sample value learning and policy extraction methods suffer from over-regularization, limiting their performance on suboptimal or compositional datasets. In this paper, we analyze key factors in in-sample learning that might potentially hinder the use of a milder constraint. We propose Actor-Critic with Temperature adjustment and In-sample Value Ensemble (ACTIVE), a novel in-sample offline RL algorithm that leverages an ensemble of V-functions for critic training and adaptively adjusts the constraint level using dual gradient descent. We theoretically show that the V-ensemble suppresses the accumulation of initial value errors, thereby mitigating overestimation. Our experiments on the D4RL benchmarks demonstrate that ACTIVE alleviates overfitting of value functions and outperforms existing in-sample methods in terms of learning stability and policy optimality.

1123. Offline RL with Smooth OOD Generalization in Convex Hull and its Neighborhood

链接：https://iclr.cc/virtual/2025/poster/28920 abstract： Offline Reinforcement Learning (RL) struggles with distributional shifts, leading to the Q-value overestimation for out-of-distribution (OOD) actions. Existing methods address this issue by imposing constraints; however, they often become overly conservative when evaluating OOD regions, which constrains the Q-function generalization. This over-constraint issue results in poor Q-value estimation and hinders policy improvement. In this paper, we introduce a novel approach to achieve better Q-value estimation by enhancing Q-function generalization in OOD regions within Convex Hull and its Neighborhood (CHN). Under the safety generalization guarantees of the CHN, we propose the Smooth Bellman Operator (SBO), which updates OOD Q-values by smoothing them with neighboring in-sample Q-values. We theoretically show that SBO approximates true Q-values for both in-sample and OOD actions within the CHN. Our practical algorithm, Smooth Q-function OOD Generalization (SQOG), empirically alleviates the over-constraint issue, achieving near-accurate Q-value estimation. On the D4RL benchmarks, SQOG outperforms existing state-of-the-art methods in both performance and computational efficiency. Code is available at .

1124. Revisiting Mode Connectivity in Neural Networks with Bezier Surface

链接：https://iclr.cc/virtual/2025/poster/31207 abstract： Understanding the loss landscapes of neural networks (NNs) is critical for optimizing model performance. Previous research has identified the phenomenon of mode connectivity on curves, where two well-trained NNs can be connected by a continuous path in parameter space where the path maintains nearly constant loss. In this work, we extend the concept of mode connectivity to explore connectivity on surfaces, significantly broadening its applicability and unlocking new opportunities. While initial attempts to connect models via linear surfaces in parameter space were unsuccessful, we propose a novel optimization technique that consistently discovers Bézier surfaces with low-loss and high-accuracy connecting multiple NNs in a nonlinear manner. We further demonstrate that even without optimization, mode connectivity exists in certain cases of Bézier surfaces, where the models are carefully selected and combined linearly. This approach provides a deeper and more comprehensive understanding of the loss landscape and offers a novel way to identify models with enhanced performance for model averaging and output ensembling. We demonstrate the effectiveness of our method on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets using VGG16, ResNet18, and ViT architectures.

1125. Inverse decision-making using neural amortized Bayesian actors

链接：https://iclr.cc/virtual/2025/poster/27641 abstract： Bayesian observer and actor models have provided normative explanations for many behavioral phenomena in perception, sensorimotor control, and other areas of cognitive science and neuroscience. They attribute behavioral variability and biases to interpretable entities such as perceptual and motor uncertainty, prior beliefs, and behavioral costs. However, when extending these models to more naturalistic tasks with continuous actions, solving the Bayesian decision-making problem is often analytically intractable. Inverse decision-making, i.e. performing inference over the parameters of such models given behavioral data, is computationally even more difficult. Therefore, researchers typically constrain their models to easily tractable components, such as Gaussian distributions or quadratic cost functions, or resort to numerical approximations. To overcome these limitations, we amortize the Bayesian actor using a neural network trained on a wide range of parameter settings in an unsupervised fashion. Using the pre-trained neural network enables performing efficient gradient-based Bayesian inference of the Bayesian actor model's parameters. We show on synthetic data that the inferred posterior distributions are in close alignment with those obtained using analytical solutions where they exist. Where no analytical solution is available, we recover posterior distributions close to the ground truth. We then show how our method allows for principled model comparison and how it can be used to disentangle factors that may lead to unidentifiabilities between priors and costs. Finally, we apply our method to empirical data from three sensorimotor tasks and compare model fits with different cost functions to show that it can explain individuals' behavioral patterns.

1126. Understanding and Mitigating Bottlenecks of State Space Models through the Lens of Recency and Over-smoothing

链接：https://iclr.cc/virtual/2025/poster/28272 abstract： Structured State Space Models (SSMs) have emerged as alternatives to transformers.While SSMs are often regarded as effective in capturing long-sequence dependencies, we rigorously demonstrate that they are inherently limited by strong recency bias.Our empirical studies also reveal that this bias impairs the models' ability to recall distant information and introduces robustness issues. Our scaling experiments then discovered that deeper structures in SSMs can facilitate the learning of long contexts.However, subsequent theoretical analysis reveals that as SSMs increase in depth, they exhibit another inevitable tendency toward over-smoothing, e.g., token representations becoming increasingly indistinguishable.This fundamental dilemma between recency and over-smoothing hinders the scalability of existing SSMs. Inspired by our theoretical findings, we propose to polarize two channels of the state transition matrices in SSMs, setting them to zero and one, respectively, simultaneously addressing recency bias and over-smoothing.Experiments demonstrate that our polarization technique consistently enhances the associative recall accuracy of long-range tokens and unlocks SSMs to benefit further from deeper architectures.All source codes are released at https://github.com/VITA-Group/SSM-Bottleneck.

1127. Approximating Full Conformal Prediction for Neural Network Regression with Gauss-Newton Influence

链接：https://iclr.cc/virtual/2025/poster/27902 abstract： Uncertainty quantification is an important prerequisite for the deployment of deep learning models in safety-critical areas. Yet, this hinges on the uncertainty estimates being useful to the extent the prediction intervals are well-calibrated and sharp. In the absence of inherent uncertainty estimates (e.g. pretrained models predicting only point estimates), popular approaches that operate post-hoc include Laplace’s method and split conformal prediction (split-CP). However, Laplace’s method can be miscalibrated when the model is misspecified and split-CP requires sample splitting, and thus comes at the expense of statistical efficiency. In this work, we construct prediction intervals for neural network regressors post-hoc without held-out data. This is achieved by approximating the full conformal prediction method (full-CP). Whilst full-CP nominally requires retraining the model for every test point and candidate label, we propose to train just once and locally perturb model parameters using Gauss-Newton influence to approximate the effect of retraining. Coupled with linearization of the network, we express the absolute residual nonconformity score as a piecewise linear function of the candidate label allowing for an efficient procedure that avoids the exhaustive search over the output space. On standard regression benchmarks and bounding box localization, we show the resulting prediction intervals are locally-adaptive and often tighter than those of split-CP.

1128. SWEb: A Large Web Dataset for the Scandinavian Languages

链接：https://iclr.cc/virtual/2025/poster/32052 abstract： This paper presents the hitherto largest pretraining dataset for the Scandinavian languages: the Scandinavian WEb (SWEb), comprising over one trillion tokens. The paper details the collection and processing pipeline, and introduces a novel model-based text extractor that significantly reduces complexity in comparison with rule-based approaches. We also introduce a new cloze-style benchmark for evaluating language models in Swedish, and use this test to compare models trained on the SWEb data to models trained on FineWeb, with competitive results. All data, models and code are shared openly.

1129. Reveal Object in Lensless Photography via Region Gaze and Amplification

链接：https://iclr.cc/virtual/2025/poster/30401 abstract： Detecting concealed objects, such as in vivo lesions or camouflage, requires customized imaging systems. Lensless cameras, being compact and flexible, offer a promising alternative to bulky lens systems. However, the absence of lenses leads to measurements lacking visual semantics, posing significant challenges for concealed object detection (COD). To tackle this issue, we propose a region gaze-amplification network (RGANet) for progressively exploiting concealed objects from lensless imaging measurements. Specifically, a region gaze module (RGM) is proposed to mine spatial-frequency cues informed by biological and psychological mechanisms, and a region amplifier (RA) is designed to amplify the details of object regions to enhance COD performance. Furthermore, we contribute the first relevant dataset as a benchmark to prosper the lensless imaging community. Extensive experiments demonstrate the exciting performance of our method. Our codes will be released in \url{https://github.com/YXJ-NTU/Lensless-COD}.

1130. Leveraging Submodule Linearity Enhances Task Arithmetic Performance in LLMs

链接：https://iclr.cc/virtual/2025/poster/28670 abstract： Task arithmetic is a straightforward yet highly effective strategy for model merging, enabling the resultant model to exhibit multi-task capabilities. Recent research indicates that models demonstrating linearity enhance the performance of task arithmetic. In contrast to existing methods that rely on the global linearization of the model, we argue that this linearity already exists within the model's submodules. In particular, we present a statistical analysis and show that submodules (e.g., layers, self-attentions, and MLPs) exhibit significantly higher linearity than the overall model. Based on these findings, we propose an innovative model merging strategy that independently merges these

submodules. Especially, we derive a closed-form solution for optimal merging weights grounded in the linear properties of these submodules. Experimental results demonstrate that our method consistently outperforms the standard task arithmetic approach and other established baselines across different model scales and various tasks. This result highlights the benefits of leveraging the linearity of submodules and provides a new perspective for exploring solutions for effective and practical multi-task model merging.

1131. ReSi: A Comprehensive Benchmark for Representational Similarity Measures

链接：https://iclr.cc/virtual/2025/poster/29754 abstract： Measuring the similarity of different representations of neural architectures is a fundamental task and an open research challenge for the machine learning community. This paper presents the first comprehensive benchmark for evaluating representational similarity measures based on well-defined groundings of similarity. The representational similarity (ReSi) benchmark consists of (i) six carefully designed tests for similarity measures, (ii) 24 similarity measures, (iii) 14 neural network architectures, and (iv) seven datasets, spanning over the graph, language, and vision domains. The benchmark opens up several important avenues of research on representational similarity that enable novel explorations and applications of neural architectures. We demonstrate the utility of the ReSi benchmark by conducting experiments on various neural network architectures, real world datasets and similarity measures. All components of the benchmark are publicly available and thereby facilitate systematic reproduction and production of research results. The benchmark is extensible, future research can build on and further expand it. We believe that the ReSi benchmark can serve as a sound platform catalyzing future research that aims to systematically evaluate existing and explore novel ways of comparing representations of neural architectures. ReSi is available at https://github.com/mklabunde/resi.

1132. Operator Deep Smoothing for Implied Volatility

链接：https://iclr.cc/virtual/2025/poster/32104 abstract： We devise a novel method for nowcasting implied volatility based on neural operators.Better known as implied volatility smoothing in the financial industry, nowcasting of implied volatility means constructing a smooth surface that is consistent with the prices presently observed on a given option market.Option price data arises highly dynamically in ever-changing spatial configurations, which poses a major limitation to foundational machine learning approaches using classical neural networks.While large models in language and image processing deliver breakthrough results on vast corpora of raw data, in financial engineering the generalization from big historical datasets has been hindered by the need for considerable data pre-processing.In particular, implied volatility smoothing has remained an instance-by-instance, hands-on process both for neural network-based and traditional parametric strategies.Our general operator deep smoothing approach, instead, directly maps observed data to smoothed surfaces.We adapt the graph neural operator architecture to do so with high accuracy on ten years of raw intraday S&P 500 options data, using a single model instance.The trained operator adheres to critical no-arbitrage constraints and is robust with respect to subsampling of inputs (occurring in practice in the context of outlier removal).We provide extensive historical benchmarks and showcase the generalization capability of our approach in a comparison with classical neural networks and SVI, an industry standard parametrization for implied volatility. The operator deep smoothing approach thus opens up the use of neural networks on large historical datasets in financial engineering.

1133. Learning from negative feedback, or positive feedback or both

链接：https://iclr.cc/virtual/2025/poster/31025 abstract： Existing preference optimization methods often assume scenarios where paired preference feedback (preferred/positive vs. dis-preferred/negative examples) is available. This requirement limits their applicability in scenarios where only unpaired feedback—for example, either positive or negative— is available. To address this, we introduce a novel approach that decouples learning from positive and negative feedback. This decoupling enables control over the influence of each feedback type and, importantly, allows learning even when only one feedback type is present. A key contribution is demonstrating stable learning from negative feedback alone, a capability not well-addressed by current methods. Our approach builds upon the probabilistic framework introduced in (Dayan and Hinton, 1997), which uses expectation-maximization (EM) to directly optimize the probability of positive outcomes (as opposed to classic expected reward maximization). We address a key limitation in current EM-based methods: they solely maximize the likelihood of positive examples, while neglecting negative ones. We show how to extend EM algorithms to explicitly incorporate negative examples, leading to a theoretically grounded algorithm that offers an intuitive and versatile way to learn from both positive and negative feedback. We evaluate our approach for training language models based on human feedback as well as training policies for sequential decision-making problems, where learned value functions are available.

1134. Vertical Federated Learning with Missing Features During Training and Inference

链接：https://iclr.cc/virtual/2025/poster/29822 abstract： Vertical federated learning trains models from feature-partitioned datasets across multiple clients, who collaborate without sharing their local data. Standard approaches assume that all feature partitions are available during both training and inference. Yet, in practice, this assumption rarely holds, as for many samples only a subset of the clients observe their partition. However, not utilizing incomplete samples during training harms generalization, and not supporting them during inference limits the utility of the model. Moreover, if any client leaves the federation after training, its partition becomes unavailable, rendering the learned model unusable. Missing feature blocks are

therefore a key challenge limiting the applicability of vertical federated learning in real-world scenarios. To address this, we propose LASER-VFL, a vertical federated learning method for efficient training and inference of split neural network-based models that is capable of handling arbitrary sets of partitions. Our approach is simple yet effective, relying on the sharing of model parameters and on task-sampling to train a family of predictors. We show that LASER-VFL achieves a $\mathcal{O}({1}/{\sqrt{T}})$ convergence rate for nonconvex objectives and, under the Polyak-Łojasiewicz inequality, it achieves linear convergence to a neighborhood of the optimum. Numerical experiments show improved performance of LASER-VFL over the baselines. Remarkably, this is the case even in the absence of missing features. For example, for CIFAR-100, we see an improvement in accuracy of 19.3\% when each of four feature blocks is observed with a probability of 0.5 and of 9.5\% when all features are observed. The code for this work is available at https://github.com/Valdeira/LASER-VFL.

1135. Exploiting Hidden Symmetry to Improve Objective Perturbation for DP Linear Learners with a Nonsmooth L1-Norm

链接：https://iclr.cc/virtual/2025/poster/30128 abstract： Objective Perturbation (OP) is a classic approach to differentially private (DP) convex optimization with smooth loss functions but is less understood for nonsmooth cases. In this work, we study how to apply OP to DP linear learners under loss functions with an implicit ℓ_1-norm structure, such as $\max(0,x)$ as a motivating example. We propose to first smooth out the implicit ℓ_1-norm by convolution, and then invoke standard OP. Convolution has many advantages that distinguish itself from Moreau Envelope, such as approximating from above and a higher degree of hyperparameters. These advantages, in conjunction with the symmetry of ℓ_1-norm, result in tighter pointwise approximation, which further facilitates tighter analysis of generalization risks by using pointwise bounds. Under mild assumptions on groundtruth distributions, the proposed OP-based algorithm is found to be rate-optimal, and can achieve the excess generalization risk $\mathcal{O}(\frac{1}{\sqrt{n}}+\frac{\sqrt{d\ln(1/\delta)}}{n\varepsilon})$. Experiments demonstrate the competitive performance of the proposed method to Noisy-SGD.

1136. Deep Weight Factorization: Sparse Learning Through the Lens of Artificial Symmetries

链接：https://iclr.cc/virtual/2025/poster/27917 abstract： Sparse regularization techniques are well-established in machine learning, yet their application in neural networks remains challenging due to the non-differentiability of penalties like the L_1 norm, which is incompatible with stochastic gradient descent. A promising alternative is shallow weight factorization, where weights are decomposed into two factors, allowing for smooth optimization of L_1-penalized neural networks by adding differentiable L_2 regularization to the factors. In this work, we introduce deep weight factorization, extending previous shallow approaches to more than two factors. We theoretically establish equivalence of our deep factorization with non-convex sparse regularization and analyze its impact on training dynamics and optimization. Due to the limitations posed by standard training practices, we propose a tailored initialization scheme and identify important learning rate requirements necessary for training factorized networks.We demonstrate the effectiveness of our deep weight factorization through experiments on various architectures and datasets, consistently outperforming its shallow counterpart and widely used pruning methods.

1137. The Power of LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions

链接：https://iclr.cc/virtual/2025/poster/27816 abstract： Stance detection holds great potential to improve online political discussions through its deployment in discussion platforms for purposes such as content moderation, topic summarisation or to facilitate more balanced discussions. Typically, transformer-based models are employed directly for stance detection, requiring vast amounts of data. However, the wide variety of debate topics in online political discussions makes data collection particularly challenging. LLMs have revived stance detection, but their online deployment in online political discussions faces challenges like inconsistent outputs, biases, and vulnerability to adversarial attacks. We show how LLM-generated synthetic data can improve stance detection for online political discussions by using reliable traditional stance detection models for online deployment, while leveraging the text generation capabilities of LLMs for synthetic data generation in a secure offline environment. To achieve this, (i) we generate synthetic data for specific debate questions by prompting a Mistral-7B model and show that fine-tuning with the generated synthetic data can substantially improve the performance of stance detection, while remaining interpretable and aligned with real world data. (ii) Using the synthetic data as a reference, we can improve performance even further by identifying the most informative samples in an unlabelled dataset, i.e., those samples which the stance detection model is most uncertain about and can benefit from the most. By fine-tuning with both synthetic data and the most informative samples, we surpass the performance of the baseline model that is fine-tuned on all true labels, while labelling considerably less data.

1138. A Second-Order Perspective on Model Compositionality and Incremental Learning

链接：https://iclr.cc/virtual/2025/poster/29821 abstract： The fine-tuning of deep pre-trained models has revealed compositional properties, with multiple specialized modules that can be arbitrarily composed into a single, multi-task model. However, identifying the conditions that promote compositionality remains an open issue, with recent efforts concentrating mainly on linearized networks. We conduct a theoretical study that attempts to demystify compositionality in standard non-linear networks through the second-order Taylor approximation of the loss function. The proposed formulation highlights the importance

of staying within the pre-training basin to achieve composable modules. Moreover, it provides the basis for two dual incremental training algorithms: the one from the perspective of multiple models trained individually, while the other aims to optimize the composed model as a whole. We probe their application in incremental classification tasks and highlight some valuable skills. In fact, the pool of incrementally learned modules not only supports the creation of an effective multi-task model but also enables unlearning and specialization in certain tasks. Code available at https://github.com/aimagelab/mammoth

1139. Going Beyond Feature Similarity: Effective Dataset distillation based on Class-aware Conditional Mutual Information

链接：https://iclr.cc/virtual/2025/poster/31237 abstract： Dataset distillation (DD) aims to minimize the time and memory consumption needed for training deep neural networks on large datasets, by creating a smaller synthetic dataset that has similar performance to that of the full real dataset. However, current dataset distillation methods often result in synthetic datasets that are excessively difficult for networks to learn from, due to the compression of a substantial amount of information from the original data through metrics measuring feature similarity, e,g., distribution matching (DM). In this work, we introduce conditional mutual information (CMI) to assess the class-aware complexity of a dataset and propose a novel method by minimizing CMI. Specifically, we minimize the distillation loss while constraining the class-aware complexity of the synthetic dataset by minimizing its empirical CMI from the feature space of pre-trained networks, simultaneously. Conducting on a thorough set of experiments, we show that our method can serve as a general regularization method to existing DD methods and improve the performance and training efficiency.

1140. Adversarially Robust Out-of-Distribution Detection Using Lyapunov-Stabilized Embeddings

链接：https://iclr.cc/virtual/2025/poster/30255 abstract： Despite significant advancements in out-of-distribution (OOD) detection, existing methods still struggle to maintain robustness against adversarial attacks, compromising their reliability in critical real-world applications. Previous studies have attempted to address this challenge by exposing detectors to auxiliary OOD datasets alongside adversarial training. However, the increased data complexity inherent in adversarial training, and the myriad of ways that OOD samples can arise during testing, often prevent these approaches from establishing robust decision boundaries. To address these limitations, we propose AROS, a novel approach leveraging neural ordinary differential equations (NODEs) with Lyapunov stability theorem in order to obtain robust embeddings for OOD detection. By incorporating a tailored loss function, we apply Lyapunov stability theory to ensure that both in-distribution (ID) and OOD data converge to stable equilibrium points within the dynamical system. This approach encourages any perturbed input to return to its stable equilibrium, thereby enhancing the model’s robustness against adversarial perturbations. To not use additional data, we generate fake OOD embeddings by sampling from low-likelihood regions of the ID data feature space, approximating the boundaries where OOD data are likely to reside. To then further enhance robustness, we propose the use of an orthogonal binary layer following the stable feature space, which maximizes the separation between the equilibrium points of ID and OOD samples. We validate our method through extensive experiments across several benchmarks, demonstrating superior performance, particularly under adversarial attacks. Notably, our approach improves robust detection performance from 37.8% to 80.1% on CIFAR-10 vs. CIFAR-100 and from 29.0% to 67.0% on CIFAR-100 vs. CIFAR-10. Code and pre-trained models are available at https://github.com/AdaptiveMotorControlLab/AROS.

1141. Robust Gymnasium: A Unified Modular Benchmark for Robust Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31105 abstract： Driven by inherent uncertainty and the sim-to-real gap, robust reinforcement learning (RL) seeks to improve resilience against the complexity and variability in agent-environment sequential interactions. Despite the existence of a large number of RL benchmarks, there is a lack of standardized benchmarks for robust RL. Current robust RL policies often focus on a specific type of uncertainty and are evaluated in distinct, one-off environments. In this work, we introduce Robust-Gymnasium, a unified modular benchmark designed for robust RL that supports a wide variety of disruptions across all key RL components—agents' observed state and reward, agents' actions, and the environment. Offering over sixty diverse task environments spanning control and robotics, safe RL, and multi-agent RL, it provides an open-source and user-friendly tool for the community to assess current methods and foster the development of robust RL algorithms. In addition, we benchmark existing standard and robust RL algorithms within this framework, uncovering significant deficiencies in each and offering new insights.

1142. Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF

链接：https://iclr.cc/virtual/2025/poster/29608 abstract： Reinforcement learning from human feedback (RLHF) has demonstrated great promise in aligning large language models (LLMs) with human preference. Depending on the availability of preference data, both online and offline RLHF are active areas of investigation. A key bottleneck is understanding how to incorporate uncertainty estimation in the reward function learned from the preference data for RLHF, regardless of how the preference data is collected. While the principles of optimism or pessimism under uncertainty are well-established in standard reinforcement learning (RL), a practically-implementable and theoretically-grounded form amenable to large language models is

not yet available, as standard techniques for constructing confidence intervals become intractable under arbitrary policy parameterizations.In this paper, we introduce a unified approach to online and offline RLHF --- value-incentivized preference optimization (VPO) --- which regularizes the maximum-likelihood estimate of the reward function with the corresponding value function, modulated by a sign to indicate whether the optimism or pessimism is chosen. VPO also directly optimizes the policy with implicit reward modeling, and therefore shares a simpler RLHF pipeline similar to direct preference optimization. Theoretical guarantees of VPO are provided for both online and offline settings, matching the rates of their standard RL counterparts. Moreover, experiments on text summarization, dialogue, and standard benchmarks verify the practicality and effectiveness of VPO.

1143. Self-supervised contrastive learning performs non-linear system identification

链接：https://iclr.cc/virtual/2025/poster/29828 abstract： Self-supervised learning (SSL) approaches have brought tremendous success across many tasks and domains. It has been argued that these successes can be attributed to a link between SSL and identifiable representation learning: Temporal structure and auxiliary variables ensure that latent representations are related to the true underlying generative factors of the data. Here, we deepen this connection and show that SSL can perform system identification in latent space. We propose DynCL, a framework to uncover linear, switching linear and non-linear dynamics under a non-linear observation model, give theoretical guarantees and validate them empirically.

1144. Learning on One Mode: Addressing Multi-modality in Offline Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/27958 abstract： Offline reinforcement learning (RL) seeks to learn optimal policies from static datasets without interacting with the environment. A common challenge is handling multi-modal action distributions, where multiple behaviours are represented in the data. Existing methods often assume unimodal behaviour policies, leading to suboptimal performance when this assumption is violated. We propose weighted imitation Learning on One Mode (LOM), a novel approach that focuses on learning from a single, promising mode of the behaviour policy. By using a Gaussian mixture model to identify modes and selecting the best mode based on expected returns, LOM avoids the pitfalls of averaging over conflicting actions. Theoretically, we show that LOM improves performance while maintaining simplicity in policy learning. Empirically, LOM outperforms existing methods on standard D4RL benchmarks and demonstrates its effectiveness in complex, multi-modal scenarios.

1145. Continual Slow-and-Fast Adaptation of Latent Neural Dynamics (CoSFan): Meta-Learning What-How & When to Adapt

链接：https://iclr.cc/virtual/2025/poster/30443 abstract： An increasing interest in learning to forecast for time-series of high-dimensional observations is the ability to adapt to systems with diverse underlying dynamics. Access to observations that define a stationary distribution of these systems is often unattainable, as the underlying dynamics may change over time. Naively training or retraining models at each shift may lead to catastrophic forgetting about previously-seen systems. We present a new continual meta-learning (CML) framework to realize continual slow-and fast adaptation of latent dynamics (CoSFan). We leverage a feed-forward meta-model to infer what the current system is and how to adapt a latent dynamics function to it, enabling fast adaptation to specific dynamics. We then develop novel strategies to automatically detect when a shift of data distribution occurs, with which to identify its underlying dynamics and its relation with previously-seen dynamics. In combination with fixed-memory experience replay mechanisms, this enables continual slow update of the what-how meta-model. Empirical studies demonstrated that both the meta- and continual-learning component was critical for learning to forecast across non-stationary distributions of diverse dynamics systems, and the feed-forward meta-model combined with task-aware/-relational continual learning strategies significantly outperformed existing CML alternatives.

1146. Reflective Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/27783 abstract： Novel view synthesis has experienced significant advancements owing to increasingly capable NeRF- and 3DGS-based methods. However, reflective object reconstruction remains challenging, lacking a proper solution to achieve real-time, high-quality rendering while accommodating inter-reflection. To fill this gap, we introduce a Reflective Gaussian splatting (Ref-Gaussian) framework characterized with two components: (I) Physically based deferred rendering that empowers the rendering equation with pixel-level material properties via formulating split-sum approximation; (II) Gaussian-grounded inter-reflection that realizes the desired inter-reflection function within a Gaussian splatting paradigm for the first time. To enhance geometry modeling, we further introduce material-aware normal propagation and an initial per-Gaussian shading stage, along with 2D Gaussian primitives. Extensive experiments on standard datasets demonstrate that Ref-Gaussian surpasses existing approaches in terms of quantitative metrics, visual quality, and compute efficiency. Further, we show that our method serves as a unified solution for both reflective and non-reflective scenes, going beyond the previous alternatives focusing on only reflective scenes. Also, we illustrate that Ref-Gaussian supports more applications such as relighting and editing.

1147. Proxy Denoising for Source-Free Domain Adaptation

链接：https://iclr.cc/virtual/2025/poster/30349 abstract： Source-Free Domain Adaptation (SFDA) aims to adapt a pre-trained source model to an unlabeled target domain with no access to the source data. Inspired by the success of large Vision-Language (ViL) models in many applications, the latest research has validated ViL's benefit for SFDA by using their predictions as pseudo supervision. However, we observe that ViL's supervision could be noisy and inaccurate at an unknown rate, potentially introducing additional negative effects during adaption. To address this thus-far ignored challenge, we introduce a novel Proxy Denoising (ProDe) approach. The key idea is to leverage the ViL model as a proxy to facilitate the adaptation process towards the latent domain-invariant space. Concretely, we design a proxy denoising mechanism to correct ViL's predictions. This is grounded on a proxy confidence theory that models the dynamic effect of proxy's divergence against the domain-invariant space during adaptation. To capitalize the corrected proxy, we further derive a mutual knowledge distilling regularization. Extensive experiments show that ProDe significantly outperforms the current state-of-the-art alternatives under both conventional closed-set setting and the more challenging open-set, partial-set, generalized SFDA, multi-target, multi-source, and test-time settings. Our code and data are available at https://github.com/tntek/source-free-domain-adaptation.

1148. GeoLoRA: Geometric integration for parameter efficient fine-tuning

链接：https://iclr.cc/virtual/2025/poster/29084 abstract： Low-Rank Adaptation (LoRA) has become a widely used method for parameter-efficient fine-tuning of large-scale, pre-trained neural networks. However, LoRA and its extensions face several challenges, including the need for rank adaptivity, robustness, and computational efficiency during the fine-tuning process. We introduce GeoLoRA, a novel approach that addresses these limitations by leveraging dynamical low-rank approximation theory. GeoLoRA requires only a single backpropagation pass over the small-rank adapters, significantly reducing computational cost as compared to similar dynamical low-rank training methods and making it faster than popular baselines such as AdaLoRA. This allows GeoLoRA to efficiently adapt the allocated parameter budget across the model, achieving smaller low-rank adapters compared to heuristic methods like AdaLoRA and LoRA, while maintaining critical convergence, descent, and error-bound theoretical guarantees. The resulting method is not only more efficient but also more robust to varying hyperparameter settings. We demonstrate the effectiveness of GeoLoRA on several state-of-the-art benchmarks, showing that it outperforms existing methods in bothaccuracy and computational efficiency

1149. Concept Bottleneck Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29660 abstract： We introduce Concept Bottleneck Large Language Models (CB-LLMs), a novel framework for building inherently interpretable Large Language Models (LLMs). In contrast to traditional black-box LLMs that rely on limited post-hoc interpretations, CB-LLMs integrate intrinsic interpretability directly into the LLMs -- allowing accurate explanations with scalability and transparency. We build CB-LLMs for two essential NLP tasks: text classification and text generation. In text classification, CB-LLMs is competitive with, and at times outperforms, traditional black-box models while providing explicit and interpretable reasoning. For the more challenging task of text generation, interpretable neurons in CB-LLMs enable precise concept detection, controlled generation, and safer outputs. The embedded interpretability empowers users to transparently identify harmful content, steer model behavior, and unlearn undesired concepts -- significantly enhancing the safety, reliability, and trustworthiness of LLMs, which are critical capabilities notably absent in existing language models.

1150. A transfer learning framework for weak to strong generalization

链接：https://iclr.cc/virtual/2025/poster/29741 abstract： Modern large language model (LLM) alignment techniques rely on human feedback, but it is unclear whether the techniques fundamentally limit the capabilities of aligned LLMs. In particular, it is unclear whether it is possible to align (stronger) LLMs with superhuman capabilities with (weaker) human feedback without degrading their capabilities. This is an instance of the weak-to-strong generalization problem: using weaker (less capable) feedback to train a stronger (more capable) model. We prove that weak-to-strong generalization is possible by eliciting latent knowledge from pre-trained LLMs. In particular, we cast the weak-to-strong generalization problem as a transfer learning problem in which we wish to transfer a latent concept from a weak model to a strong pre-trained model. We prove that a naive fine-tuning approach suffers from fundamental limitations, but an alternative refinement-based approach suggested by the problem structure provably overcomes the limitations of fine-tuning. Finally, we demonstrate the practical applicability of the refinement approach in multiple LLM alignment tasks.

1151. Higher-Order Graphon Neural Networks: Approximation and Cut Distance

链接：https://iclr.cc/virtual/2025/poster/29581 abstract： Graph limit models, like *graphons* for limits of dense graphs, have recently been used to study size transferability of graph neural networks (GNNs). While most literature focuses on message passing GNNs (MPNNs), in this work we attend to the more powerful *higher-order* GNNs. First, we extend the k-WL test for graphons (Böker, 2023) to the graphon-signal space and introduce *signal-weighted homomorphism densities* as a key tool. As an exemplary focus, we generalize *Invariant Graph Networks* (IGNs) to graphons, proposing *Invariant Graphon Networks* (IWNs) defined via a subset of the IGN basis corresponding to bounded linear operators. Even with this restricted basis, we show that IWNs of order k are at least as powerful as the k-WL test, and we establish universal approximation results for graphon-signals in L^p distances. This significantly extends the prior work of Cai & Wang (2022), showing that IWNs—a subset of their *IGN-small*—retain effectively the same expressivity as the full IGN basis in the limit. In contrast to their approach, our blueprint of

IWNs also aligns better with the geometry of graphon space, for example facilitating comparability to MPNNs. We highlight that, while typical higher-order GNNs are discontinuous w.r.t.\ cut distance—which causes their lack of convergence and is inherently tied to the definition of k-WL—their transferability remains comparable to MPNNs.

1152. Generalization, Expressivity, and Universality of Graph Neural Networks on Attributed Graphs

链接：https://iclr.cc/virtual/2025/poster/28255 abstract： We analyze the universality and generalization of graph neural networks (GNNs) on attributed graphs, i.e., with node attributes. To this end, we propose pseudometrics over the space of all attributed graphs that describe the fine-grained expressivity of GNNs. Namely, GNNs are both Lipschitz continuous with respect to our pseudometrics and can separate attributed graphs that are distant in the metric. Moreover, we prove that the space of all attributed graphs is relatively compact with respect to our metrics. Based on these properties, we prove a universal approximation theorem for GNNs and generalization bounds for GNNs on any data distribution of attributed graphs. The proposed metrics compute the similarity between the structures of attributed graphs via a hierarchical optimal transport between computation trees. Our work extends and unites previous approaches which either derived theory only for graphs with no attributes, derived compact metrics under which GNNs are continuous but without separation power, or derived metrics under which GNNs are continuous and separate points but the space of graphs is not relatively compact, which prevents universal approximation and generalization analysis.

1153. LiveXiv - A Multi-Modal live benchmark based on Arxiv papers content

链接：https://iclr.cc/virtual/2025/poster/29572 abstract： The large-scale training of multi-modal models on data scraped from the web has shown outstanding utility in infusing these models with the required world knowledge to perform effectively on multiple downstream tasks. However, one downside of scraping data from the web can be the potential sacrifice of the benchmarks on which the abilities of these models are often evaluated. To safeguard against test data contamination and to truly test the abilities of these foundation models we propose LiveXiv: A scalable evolving live benchmark based on scientific ArXiv papers. LiveXiv accesses domain-specific manuscripts at any given timestamp and proposes to automatically generate visual question-answer pairs (VQA). This is done without any human-in-the-loop, using the multi-modal content in the manuscripts, like graphs, charts, and tables. Moreover, we introduce an efficient evaluation approach that estimates the performance of all models on the evolving benchmark using evaluations of only a subset of models. This significantly reduces the overall evaluation cost. We benchmark multiple open and proprietary Large Multi-modal Models (LMMs) on the first version of our benchmark, showing its challenging nature and exposing the models' true abilities, avoiding contamination. Lastly, in our commitment to high quality, we have collected and evaluated a manually verified subset. By comparing its overall results to our automatic annotations, we have found that the performance variance is indeed minimal (<2.5%). Our dataset is available online on HuggingFace, and our code is available here.

1154. Generalization Bounds for Canonicalization: A Comparative Study with Group Averaging

链接：https://iclr.cc/virtual/2025/poster/28434 abstract： Canonicalization, a popular method for generating invariant or equivariant function classes from arbitrary function sets, involves initial data projection onto a reduced input space subset, followed by applying any learning method to the projected dataset. Despite recent research on the expressive power and continuity of functions represented by canonicalization, its generalization capabilities remain less explored. This paper addresses this gap by theoretically examining the generalization benefits and sample complexity of canonicalization, comparing them with group averaging, another popular technique for creating invariant or equivariant function classes. Our findings reveal two distinct regimes where canonicalization may outperform or underperform compared to group averaging, with precise quantification of this phase transition in terms of sample size, group action characteristics, and a newly introduced concept of alignment.To the best of our knowledge, this study represents the first theoretical exploration of such behavior, offering insights into the relative effectiveness of canonicalization and group averaging under varying conditions.

1155. Scalable Influence and Fact Tracing for Large Language Model Pretraining

链接：https://iclr.cc/virtual/2025/poster/28824 abstract： Training data attribution (TDA) methods aim to attribute model outputs back to specific training examples, and the application of these methods to large language model (LLM) outputs could significantly advance model transparency and data curation. However, it has been challenging to date to apply these methods to the full scale of LLM pretraining. In this paper, we refine existing gradient-based methods to work effectively at scale, allowing us to retrieve influential examples for an 8B-parameter language model from a pretraining corpus of over 160B tokens with no need for subsampling or pre-filtering. Our method combines several techniques, including optimizer state correction, a task-specific Hessian approximation, and normalized encodings, which we find to be critical for performance at scale. In quantitative evaluations on a fact tracing task, our method performs best at identifying examples that influence model predictions, but classical, model-agnostic retrieval methods such as BM25 still perform better at finding passages which explicitly contain relevant facts. These results demonstrate a misalignment between factual attribution and causal influence. With increasing model size and training tokens, we find that influence more closely aligns with factual attribution. Finally, we examine different

types of examples identified as influential by our method, finding that while many directly entail a particular fact, others support the same output by reinforcing priors on relation types, common entities, and names. We release our prompt set and model outputs, along with a web-based visualization tool to explore influential examples for factual predictions, commonsense reasoning, arithmetic, and open-ended generation for an 8B-parameter LLM.

1156. Open-Set Graph Anomaly Detection via Normal Structure Regularisation

链接：https://iclr.cc/virtual/2025/poster/28580 abstract： This paper considers an important Graph Anomaly Detection (GAD) task, namely open-set GAD, which aims to train a detection model using a small number of normal and anomaly nodes (referred to as *seen anomalies*) to detect both seen anomalies and *unseen anomalies* (*i.e.*, anomalies that cannot be illustrated the training anomalies). Those labelled training data provide crucial prior knowledge about abnormalities for GAD models, enabling substantially reduced detection errors. However, current supervised GAD methods tend to over-emphasise fitting the seen anomalies, leading to many errors of detecting the unseen anomalies as normal nodes. Further, existing open-set AD models were introduced to handle Euclidean data, failing to effectively capture discriminative features from graph structure and node attributes for GAD. In this work, we propose a novel open-set GAD approach, namely $\underline{n}ormal$ $\underline{s}tructure$ $\underline{reg}ularisation$ (**NSReg**), to achieve generalised detection ability to unseen anomalies, while maintaining its effectiveness on detecting seen anomalies. The key idea in NSReg is to introduce a regularisation term that enforces the learning of compact, semantically-rich representations of normal nodes based on their structural relations to other nodes. When being optimised with supervised anomaly detection losses, the regularisation term helps incorporate strong normality into the modelling, and thus, it effectively avoids over-fitting the seen anomalies and learns a better normality decision boundary, largely reducing the false negatives of detecting unseen anomalies as normal. Extensive empirical results on seven real-world datasets show that NSReg significantly outperforms state-of-the-art competing methods by at least 14% AUC-ROC on the unseen anomaly classes and by 10% AUC-ROC on all anomaly classes. Code and datasets are available at https://github.com/mala-lab/NSReg.

1157. Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax

链接：https://iclr.cc/virtual/2025/poster/28477 abstract： Deep InfoMax (DIM) is a well-established method for self-supervised representation learning (SSRL) based on maximization of the mutual information between the input and the output of a deep neural network encoder. Despite the DIM and contrastive SSRL in general being well-explored, the task of learning representations conforming to a specific distribution (i.e., distribution matching, DM) is still under-addressed. Motivated by the importance of DM to several downstream tasks (including generative modeling, disentanglement, outliers detection and other), we enhance DIM to enable automatic matching of learned representations to a selected prior distribution. To achieve this, we propose injecting an independent noise into the normalized outputs of the encoder, while keeping the same InfoMax training objective. We show that such modification allows for learning uniformly and normally distributed representations, as well as representations of other absolutely continuous distributions. Our approach is tested on various downstream tasks. The results indicate a moderate trade-off between the performance on the downstream tasks and quality of DM.

1158. DreamCatalyst: Fast and High-Quality 3D Editing via Controlling Editability and Identity Preservation

链接：https://iclr.cc/virtual/2025/poster/30362 abstract： Score distillation sampling (SDS) has emerged as an effective framework in text-driven 3D editing tasks, leveraging diffusion models for 3D-consistent editing. However, existing SDS-based 3D editing methods suffer from long training times and produce low-quality results. We identify that the root cause of this performance degradation is their conflict with the sampling dynamics of diffusion models. Addressing this conflict allows us to treat SDS as a diffusion reverse process for 3D editing via sampling from data space. In contrast, existing methods naively distill the score function using diffusion models. From these insights, we propose DreamCatalyst, a novel framework that considers these sampling dynamics in the SDS framework. Specifically, we devise the optimization process of our DreamCatalyst to approximate the diffusion reverse process in editing tasks, thereby aligning with diffusion sampling dynamics. As a result, DreamCatalyst successfully reduces training time and improves editing quality. Our method offers two modes: (1) a fast mode that edits Neural Radiance Fields (NeRF) scenes approximately 23 times faster than current state-of-the-art NeRF editing methods, and (2) a high-quality mode that produces superior results about 8 times faster than these methods. Notably, our high-quality mode outperforms current state-of-the-art NeRF editing methods in terms of both speed and quality. DreamCatalyst also surpasses the state-of-the-art 3D Gaussian Splatting (3DGS) editing methods, establishing itself as an effective and model-agnostic 3D editing solution.

1159. Enhance Multi-View Classification Through Multi-Scale Alignment and Expanded Boundary

链接：https://iclr.cc/virtual/2025/poster/28081 abstract： Multi-view classification aims at unifying the data from multiple views to complementarily enhance the classification performance. Unfortunately, two major problems in multi-view data are damaging model performance. The first is feature heterogeneity, which makes it hard to fuse features from different views. Considering

this, we introduce a multi-scale alignment module, including an instance-scale alignment module and a prototype-scale alignment module to mine the commonality from an inter-view perspective and an inter-class perspective respectively, jointly alleviating feature heterogeneity. The second is information redundancy which easily incurs ambiguous data to blur class boundaries and impair model generalization. Therefore, we propose a novel expanded boundary by extending the original class boundary with fuzzy set theory, which adaptively adjusts the boundary to fit ambiguous data. By integrating the expanded boundary into the prototype-scale alignment module, our model further tightens the produced representations and reduces boundary ambiguity. Additionally, compared with the original class boundary, the expanded boundary preserves more margins for classifying unseen data, which guarantees the model generalization. Extensive experiment results across various real-world datasets demonstrate the superiority of the proposed model against existing state-of-the-art methods.

1160. OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer

链接：https://iclr.cc/virtual/2025/poster/30293 abstract： Open-vocabulary multiple object tracking aims to generalize trackers to unseen categories during training, enabling their application across a variety of real-world scenarios. However, the existing open-vocabulary tracker is constrained by its framework structure, isolated frame-level perception, and insufficient modal interactions, which hinder its performance in open-vocabulary classification and tracking. In this paper, we propose OVTR (End-to-End Open-Vocabulary Multiple Object Tracking with TRansformer), the first end-to-end open-vocabulary tracker that models motion, appearance, and category simultaneously. To achieve stable classification and continuous tracking, we design the CIP (Category Information Propagation) strategy, which establishes multiple high-level category information priors for subsequent frames. Additionally, we introduce a dual-branch structure for generalization capability and deep multimodal interaction, and incorporate protective strategies in the decoder to enhance performance. Experimental results show that our method surpasses previous trackers on the open-vocabulary MOT benchmark while also achieving faster inference speeds and significantly reducing preprocessing requirements. Moreover, the experiment transferring the model to another dataset demonstrates its strong adaptability.

1161. Variational Best-of-N Alignment

链接：https://iclr.cc/virtual/2025/poster/29381 abstract： Best-of-N (BoN) is a popular and effective algorithm for aligning language models to human preferences. The algorithm works as follows: at inference time, N samples are drawn from the language model, and the sample with the highest reward, as judged by a reward model, is returned as the output. Despite its effectiveness, BoN is computationally expensive; it reduces sampling throughput by a factor of N. To make BoN more efficient at inference time, one strategy is to fine-tune the language model to mimic what BoN does during inference. To achieve this, we derive the distribution induced by the BoN algorithm. We then propose to fine-tune the language model to minimize backward KL divergence to the BoN distribution. Our approach is analogous to mean-field variational inference and, thus, we term it variational BoN (vBoN). To the extent this fine-tuning is successful and we end up with a good approximation, we have reduced the inference cost by a factor of N. Our experiments on controlled generation and summarization tasks show that BoN is the most effective alignment method, and our variational approximation to BoN achieves the closest performance to BoN and surpasses models fine-tuned using the standard KL-constrained RL objective. In the controlled generation task, vBoN appears more frequently on the Pareto frontier of reward and KL divergence compared to other alignment methods. In the summarization task, vBoN achieves high reward values across various sampling temperatures.

1162. Learning Efficient Positional Encodings with Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30628 abstract： Positional encodings (PEs) are essential for effective graph representation learning because they provide position awareness in inherently position-agnostic transformer architectures and increase the expressive capacity of Graph Neural Networks (GNNs). However, designing powerful and efficient PEs for graphs poses significant challenges due to the absence of canonical node ordering and the scale of the graph. In this work, we identify four key properties that graph PEs should satisfy: stability, expressive power, scalability, and genericness. We find that existing eigenvector-based PE methods often fall short of jointly satisfying these criteria. To address this gap, we introduce PEARL, a novel framework of learnable PEs for graphs. Our primary insight is that message-passing GNNs function as nonlinear mappings of eigenvectors, enabling the design of GNN architectures for generating powerful and efficient PEs. A crucial challenge lies in initializing node features in a manner that is both expressive and permutation equivariant. We tackle this by initializing GNNs with random node inputs or standard basis vectors, thereby unlocking the expressive power of message-passing operations, while employing statistical pooling functions to maintain permutation equivariance. Our analysis demonstrates that PEARL approximates equivariant functions of eigenvectors with linear complexity, while rigorously establishing its stability and high expressive power. Experimental evaluations show that PEARL outperforms lightweight versions of eigenvector-based PEs and achieves comparable performance to full eigenvector-based PEs, but with one or two orders of magnitude lower complexity. Our code is available at https://github.com/ehejin/Pearl-PE.

1163. Training Neural Networks as Recognizers of Formal Languages

链接：https://iclr.cc/virtual/2025/poster/29163 abstract： Characterizing the computational power of neural network architectures in terms of formal language theory remains a crucial line of research, as it describes lower and upper bounds on the reasoning capabilities of modern AI. However, when empirically testing these bounds, existing work often leaves a discrepancy between experiments and the formal claims they are meant to support. The problem is that formal language theory

pertains specifically to recognizers: machines that receive a string as input and classify whether it belongs to a language. On the other hand, it is common instead to evaluate language models on proxy tasks, e.g., language modeling or sequence-to-sequence transduction, that are similar in only an informal sense to the underlying theory. We correct this mismatch by training and evaluating neural networks directly as binary classifiers of strings, using a general method that can be applied to a wide variety of languages. As part of this, we extend an algorithm recently proposed by Snæbjarnarson et al. (2025) for efficient length-controlled sampling of strings from regular languages. We provide results on a variety of languages across the Chomsky hierarchy for three neural architectures: a simple RNN, an LSTM, and a causally-masked transformer. We find that the RNN and LSTM often outperform the transformer, and that auxiliary training objectives such as language modeling can help, although no single objective uniformly improves performance across languages and architectures. Our contributions will facilitate theoretically sound empirical testing of language recognition claims in future work. We have released our datasets as a benchmark called FLaRe (Formal Language Recognition), along with our code.

1164. Compositional simulation-based inference for time series

链接：https://iclr.cc/virtual/2025/poster/27993 abstract： Amortized simulation-based inference (SBI) methods train neural networks on simulated data to perform Bayesian inference. While this strategy avoids the need for tractable likelihoods, it often requires a large number of simulations and has been challenging to scale to time series data. Scientific simulators frequently emulate real-world dynamics through thousands of single-state transitions over time. We propose an SBI approach that can exploit such Markovian simulators by locally identifying parameters consistent with individual state transitions. We then compose these local results to obtain a posterior over parameters that align with the entire time series observation. We focus on applying this approach to neural posterior score estimation but also show how it can be applied, e.g., to neural likelihood (ratio) estimation. We demonstrate that our approach is more simulation-efficient than directly estimating the global posterior on several synthetic benchmark tasks and simulators used in ecology and epidemiology. Finally, we validate scalability and simulation efficiency of our approach by applying it to a high-dimensional Kolmogorov flow simulator with around one million data dimensions.

1165. Controllable Context Sensitivity and the Knob Behind It

链接：https://iclr.cc/virtual/2025/poster/30159 abstract： When making predictions, a language model must trade off how much it relies on its context vs. its prior knowledge.Choosing how sensitive the model is to its context is a fundamental functionality, as it enables the model to excel at tasks like retrieval-augmented generation and question-answering. In this paper, we search for a knob which controls this sensitivity, determining whether language models answer from the context or their prior knowledge.To guide this search, we design a task for controllable context sensitivity. In this task, we first feed the model a context ("Paris is in England") and a question ("Where is Paris?"); we then instruct the model to either use its prior or contextual knowledge and evaluate whether it generates the correct answer for both intents (either "France" or "England").When fine-tuned on this task, instruct versions of Llama-3.1, Mistral-v0.3, and Gemma-2 can solve it with high accuracy (85-95%). Analyzing these high-performing models, we narrow down which layers may be important to context sensitivity using a novel linear time algorithm. Then, in each model, we identify a 1-D subspace in a single layer that encodes whether the model follows context or prior knowledge.Interestingly, while we identify this subspace in a fine-tuned model, we find that the exact same subspace serves as an effective knob in not only that model but also non-fine-tuned instruct and base models of that model family.Finally, we show a strong correlation between a model's performance and how distinctly it separates context-agreeing from context-ignoring answers in this subspace.These results suggest a single fundamental subspace facilitates how the model chooses between context and prior knowledge.

1166. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages

链接：https://iclr.cc/virtual/2025/poster/29188 abstract： Despite recent advances in multimodal large language models (MLLMs), their development has predominantly focused on English- and western-centric datasets and tasks, leaving most of the world's languages and diverse cultural contexts underrepresented. This paper introduces PANGEA, a multilingual multimodal LLM trained on PANGEAINS, a diverse 6M instruction dataset spanning 39 languages. PANGEAINS features: 1) high-quality English instructions, 2) carefully machine-translated instructions, and 3) culturally relevant multimodal tasks to ensure cross-cultural coverage. To rigorously assess models' capabilities, we introduce PANGEABENCH, a holistic evaluation suite encompassing 14 datasets covering 47 languages. Results show that PANGEA significantly outperforms existing open-source models in multilingual settings and diverse cultural contexts. Ablation studies further reveal the importance of English data proportions, language popularity, and the number of multimodal training samples on overall performance. We fully open-source our data, code, and trained checkpoints, to facilitate the development of inclusive and robust multilingual MLLMs, promoting equity and accessibility across a broader linguistic and cultural spectrum.

1167. KaSA: Knowledge-Aware Singular-Value Adaptation of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29826 abstract： The increasing sizes of large language models (LLMs) result in significant computational overhead and memory usage when adapting these models to specific tasks or domains. Various parameter-efficient fine-tuning (PEFT) methods have been devised to mitigate these challenges by training a small set of parameters for the task-specific updates of the model weights. Among PEFT methods, LoRA stands out for its simplicity and

efficiency, inspiring the development of a series of variants. However, LoRA and its successors disregard the knowledge that is noisy or irrelevant to the targeted task, detrimentally impacting model performance and leading to suboptimality. To address this limitation, we introduce Knowledge-aware Singular-value Adaptation (KaSA), a PEFT method that leverages singular value decomposition (SVD) with knowledge-aware singular values to dynamically activate knowledge based on its relevance to the task at hand. We conduct extensive experiments across a range of LLMs on tasks spanning natural language understanding (NLU), generation (NLG), instruction following, and commonsense reasoning. The experimental results demonstrate that KaSA consistently outperforms FFT and 14 popular PEFT baselines across 16 benchmarks and 4 synthetic datasets, underscoring our method's efficacy and adaptability. The source code of our method is available at https://github.com/juyongjiang/KaSA.

1168. Towards more rigorous evaluations of language models

链接：https://iclr.cc/virtual/2025/poster/31354 abstract： As language models (LMs) become increasingly sophisticated and existing benchmarks approach saturation, the need for rigorous evaluation methods grows more pressing. Many evaluations lack the statistical rigour needed to draw meaningful conclusions, leading to a potential over-confidence in results that might not hold up under scrutiny or replication. This post advocates for bringing fundamental statistical principles to language model evaluation, demonstrating how basic statistical analysis can provide more reliable insights into model capabilities and limitations.We show how to conduct this type of analysis using a recent paper as a case study. We hope this post serves as a tutorial for LM researchers aiming to enhance the rigor of their empirical evaluations.

1169. Continuous Ensemble Weather Forecasting with Diffusion models

链接：https://iclr.cc/virtual/2025/poster/28928 abstract： Weather forecasting has seen a shift in methods from numerical simulations to data-driven systems. While initial research in the area focused on deterministic forecasting, recent works have used diffusion models to produce skillful ensemble forecasts. These models are trained on a single forecasting step and rolled out autoregressively. However, they are computationally expensive and accumulate errors for high temporal resolution due to the many rollout steps. We address these limitations with Continuous Ensemble Forecasting, a novel and flexible method for sampling ensemble forecasts in diffusion models. The method can generate temporally consistent ensemble trajectories completely in parallel, with no autoregressive steps. Continuous Ensemble Forecasting can also be combined with autoregressive rollouts to yield forecasts at an arbitrary fine temporal resolution without sacrificing accuracy. We demonstrate that the method achieves competitive results for global weather forecasting with good probabilistic properties.

1170. Ada-K Routing: Boosting the Efficiency of MoE-based LLMs

链接：https://iclr.cc/virtual/2025/poster/30715 abstract： In the era of Large Language Models (LLMs), Mixture-of-Experts (MoE) architectures offer a promising approach to managing computational costs while scaling up model parameters. Conventional MoE-based LLMs typically employ static Top-K routing, which activates a fixed and equal number of experts for each token regardless of their significance within the context. In this paper, we propose a novel Ada-K routing strategy that dynamically adjusts the number of activated experts for each token, thereby improving the balance between computational efficiency and model performance. Specifically, our strategy incorporates learnable and lightweight allocator modules that decide customized expert resource allocation tailored to the contextual needs for each token. These allocators are designed to be fully pluggable, making it broadly applicable across all mainstream MoE-based LLMs. We leverage the Proximal Policy Optimization (PPO) algorithm to facilitate an end-to-end learning process for this non-differentiable decision-making framework. Extensive evaluations on four popular baseline models demonstrate that our Ada-K routing method significantly outperforms conventional Top-K routing. Compared to Top-K, our method achieves over 25% reduction in FLOPs and more than 20% inference speedup while still improving performance across various benchmarks. Moreover, the training of Ada-K is highly efficient. Even for Mixtral-8x22B, a MoE-based LLM with more than 140B parameters, the training time is limited to 8 hours. Detailed analysis shows that harder tasks, middle layers, and content words tend to activate more experts, providing valuable insights for future adaptive MoE system designs. Both the training code and model checkpoints will be publicly available.

1171. Balancing Bias in Two-sided Markets for Fair Stable Matchings

链接：https://iclr.cc/virtual/2025/poster/28228 abstract： The Balanced Stable Marriage (BSM) problem aims to find a stable matching in a two-sided market that minimizes the maximum dissatisfaction among two sides. The classical Deferred Acceptance algorithm merely produces an unfair stable marriage, providing optimal partners for one side while partially assigning pessimal partners to the other. Solving BSM is NP-hard, thwarting attempts to resolve the problem exactly. As the instance size increases in practice, recent studies have explored heuristics for finding a fair stable marriage but have not found an exact optimal solution for BSM efficiently. Nevertheless, in this paper we propose an efficient algorithm, Isorropia, that returns the exact optimal solution to practical BSM problem instances. Isorropia constructs two sets of candidate rotations from which it builds three sets of promising antichains, and performs local search on those three sets of promising antichains. Our extensive experimental study shows that Isorropia surpasses the time-efficiency of baselines that return the exact solution by up to three orders of magnitude.

1172. Multiview Equivariance Improves 3D Correspondence Understanding with Minimal Feature Finetuning

链接：https://iclr.cc/virtual/2025/poster/30515 abstract： Vision foundation models, particularly the ViT family, have revolutionized image understanding by providing rich semantic features. However, despite their success in 2D comprehension, their abilities on grasping 3D spatial relationships are still unclear.In this work, we evaluate and enhance the 3D awareness of ViT-based models. We begin by systematically assessing their ability to learn 3D equivariant features, specifically examining the consistency of semantic embeddings across different viewpoints. Our findings indicate that improved 3D equivariance leads to better performance on various downstream tasks, including pose estimation, tracking, and semantic transfer. Building on this insight, we propose a simple yet effective finetuning strategy based on 3D correspondences, which significantly enhances the 3D understanding of existing vision models. Remarkably, even finetuning on a single object for just one iteration results in substantial performance gains. Code is available on https://github.com/qq456cvb/3DCorrEnhance.

1173. Fundamental Limits of Prompt Tuning Transformers: Universality, Capacity and Efficiency

链接：https://iclr.cc/virtual/2025/poster/28645 abstract： We investigate the statistical and computational limits of prompt tuning for transformer-based foundation models. Our key contributions are that prompt tuning on *single-head* transformers with only a *single* self-attention layer: (i) is universal, and (ii) supports efficient (even almost-linear time) algorithms under the Strong Exponential Time Hypothesis (SETH). Statistically, we prove that prompt tuning on such the simplest possible transformers are universal approximators for sequence-to-sequence Lipschitz functions. In addition, we provide an exponential-in-dL and -in-$(1/\epsilon)$ lower bound on the required soft-prompt tokens for prompt tuning to memorize any dataset with 1-layer, 1-head transformers. Computationally, we identify a phase transition in the efficiency of prompt tuning, determined by the norm of the *soft-prompt-induced* keys and queries, and provide an upper bound criterion. Beyond this criterion, no sub-quadratic (efficient) algorithm for prompt tuning exists under SETH. Within this criterion, we showcase our theory by proving the existence of almost-linear time prompt tuning inference algorithms. These fundamental limits provide important necessary conditions for designing expressive and efficient prompt tuning methods for practitioners.

1174. Efficiently Democratizing Medical LLMs for 50 Languages via a Mixture of Language Family Experts

链接：https://iclr.cc/virtual/2025/poster/30111 abstract： Adapting medical Large Language Models to local languages can reduce barriers to accessing healthcare services, but data scarcity remains a significant challenge, particularly for low-resource languages. To address this, we first construct a high-quality medical dataset and conduct analysis to ensure its quality. In order to leverage the generalization capability of multilingual LLMs to efficiently scale to more resource-constrained languages, we explore the internal information flow of LLMs from a multilingual perspective using Mixture of Experts (MoE) modularity. Technically, we propose a novel MoE routing method that employs language-specific experts and cross-lingual routing. Inspired by circuit theory, our routing analysis revealed a \textit{Spread Out in the End} information flow mechanism: while earlier layers concentrate cross-lingual information flow, the later layers exhibit language-specific divergence. This insight directly led to the development of the Post-MoE architecture, which applies sparse routing only in the later layers while maintaining dense others. Experimental results demonstrate that this approach enhances the generalization of multilingual models to other languages while preserving interpretability. Finally, to efficiently scale the model to 50 languages, we introduce the concept of \textit{language family} experts, drawing on linguistic priors, which enables scaling the number of languages without adding additional parameters.

1175. KBLaM: Knowledge Base augmented Language Model

链接：https://iclr.cc/virtual/2025/poster/29174 abstract： In this paper, we propose Knowledge Base augmented Language Model (KBLAM), a new method for augmenting Large Language Models (LLMs) with external knowledge. KBLAM works with a knowledge base (KB) constructed from a corpus of documents, transforming each piece of knowledge in the KB into continuous key-value vector pairs via pre-trained sentence encoders with linear adapters andintegrating them into pre-trained LLMs via a specialized rectangular attention mechanism. Unlike Retrieval-Augmented Generation, KBLAM eliminates external retrieval modules, and unlike in-context learning, its computational overhead scales linearly with KB size rather than quadratically. Our approach enables integrating a large KB of more than 10K triples into an 8B pre-trained LLM of only 8K context window on one single A100 80GB GPU and allows for dynamic updates without model fine-tuning or retraining. Experiments demonstrate KBLAM's effectiveness in various tasks, including question-answering and open-ended reasoning, while providing interpretable insights into its use of the augmented knowledge. Code and datasets are available at https://github.com/microsoft/KBLaM/

1176. Fantastic Copyrighted Beasts and How (Not) to Generate Them

链接：https://iclr.cc/virtual/2025/poster/28846 abstract： Recent studies show that image and video generation models can be prompted to reproduce copyrighted content from their training data, raising serious legal con- cerns about copyright infringement. Copyrighted characters (e.g., Mario, Batman) present a significant challenge: at least one lawsuit has already awarded damages based on the generation of such characters. Consequently, commercial services like DALL·E have started deploying interventions. However, little research has systematically examined these problems: (1) Can users easily prompt models to generate copyrighted characters, even if it is unintentional?; (2) How effective are the existing mitigation strategies? To address these questions, we introduce a novel evaluation framework with metrics that assess both the generated image's similarity to copyrighted characters and its consistency with user intent, grounded in a set of popular copyrighted characters from

diverse studios and regions. We show that state-of-the-art image and video generation models can still generate characters even if characters' names are not explicitly mentioned, sometimes with only two generic keywords (e.g., prompting with "videogame, plumber" consistently gener- ates Nintendo's Mario character). We also introduce semi-automatic techniques to identify such keywords or descriptions that trigger character generation. Using this framework, we evaluate mitigation strategies, including prompt rewriting and new approaches we propose. Our findings reveal that common methods, such as DALL·E's prompt rewriting, are insufficient alone and require supplementary strategies like negative prompting. Our work provides empirical grounding for discussions on copyright mitigation strategies and offers actionable insights for model deployers implementing these safeguards.

1177. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding

链接：https://iclr.cc/virtual/2025/poster/32088 abstract： Understanding the physical world is a fundamental challenge in embodied AI, critical for enabling agents to perform complex tasks and operate safely in real-world environments. While Vision-Language Models (VLMs) have shown great promise in reasoning and task planning for embodied agents, their ability to comprehend physical phenomena remains extremely limited.To close this gap, we introduce PhysBench, a comprehensive benchmark designed to evaluate VLMs' physical world understanding capability across a diverse set of tasks. PhysBench contains 10,002 entries of interleaved video-image-text data, categorized into four major domains: physical object properties, physical object relationships, physical scene understanding, and physics-based dynamics, further divided into 19 subclasses and 8 distinct capability dimensions.Our extensive experiments, conducted on 75 representative VLMs, reveal that while these models excel in common-sense reasoning, they struggle with understanding the physical world---likely due to the absence of physical knowledge in their training data and the lack of embedded physical priors.To tackle the shortfall, we introduce PhysAgent, a novel framework that combines the generalization strengths of VLMs with the specialized expertise of vision models, significantly enhancing VLMs' physical understanding across a variety of tasks, including an 18.4\% improvement on GPT-4o.Furthermore, our results demonstrate that enhancing VLMs' physical world understanding capabilities can help embodied agents such as MOKA.We believe that PhysBench and PhysAgent offer valuable insights and contribute to bridging the gap between VLMs and physical world understanding. Project Page is here

1178. The Belief State Transformer

链接：https://iclr.cc/virtual/2025/poster/29526 abstract： We introduce the "Belief State Transformer", a next-token predictor that takes both a prefix and suffix as inputs, with a novel objective of predicting both the next token for the prefix and the previous token for the suffix. The Belief State Transformer effectively learns to solve challenging problems that conventional forward-only transformers struggle with, in a domain-independent fashion. Key to this success is learning a compact belief state that captures all relevant information necessary for accurate predictions.Empirical ablations show that each component of the model is essential in difficult scenarios where standard Transformers fall short. For the task of story writing with known prefixes and suffixes, our approach outperforms the Fill-in-the-Middle method for reaching known goals and demonstrates improved performance even when the goals are unknown. Altogether, the Belief State Transformer enables more efficient goal-conditioned decoding, better test-time inference, and high-quality text representations on small scale problems. Website: https://edwhu.github.io/bst-website

1179. CubeDiff: Repurposing Diffusion-Based Image Models for Panorama Generation

链接：https://iclr.cc/virtual/2025/poster/29958 abstract： We introduce a novel method for generating 360° panoramas from text prompts or images. Our approach leverages recent advances in 3D generation by employing multi-view diffusion models to jointly synthesize the six faces of a cubemap. Unlike previous methods that rely on processing equirectangular projections or autoregressive generation, our method treats each face as a standard perspective image, simplifying the generation process and enabling the use of existing multi-view diffusion models. We demonstrate that these models can be adapted to produce high-quality cubemaps without requiring correspondence-aware attention layers. Our model allows for fine-grained text control, generates high resolution panorama images and generalizes well beyond its training set, whilst achieving state-of-the-art results, both qualitatively and quantitatively.

1180. DRoP: Distributionally Robust Data Pruning

链接：https://iclr.cc/virtual/2025/poster/28840 abstract： In the era of exceptionally data-hungry models, careful selection of the training data is essential to mitigate the extensive costs of deep learning. Data pruning offers a solution by removing redundant or uninformative samples from the dataset, which yields faster convergence and improved neural scaling laws. However, little is known about its impact on classification bias of the trained models. We conduct the first systematic study of this effect and reveal that existing data pruning algorithms can produce highly biased classifiers. We present theoretical analysis of the classification risk in a mixture of Gaussians to argue that choosing appropriate class pruning ratios, coupled with random pruning within classes has potential to improve worst-class performance. We thus propose DRoP, a distributionally robust approach to pruning and empirically demonstrate its performance on standard computer vision benchmarks. In sharp contrast to existing algorithms, our proposed method continues improving distributional robustness at a tolerable drop of average performance as we prune more from the datasets.

1181. Improving Reasoning Performance in Large Language Models via Representation Engineering

链接：https://iclr.cc/virtual/2025/poster/30146 abstract： Recent advancements in large language models (LLMs) have resulted in increasingly anthropomorphic language concerning the ability of LLMs to reason. Whether \textit{reasoning} in LLMs should be understood to be inherently different is, however, widely debated. We propose utilizing a representation engineering approach wherein model activations are read from the residual stream of an LLM when processing a reasoning task. The activations are used to derive a control vector that is applied to the model as an inference-time intervention, modulating the representational space of the model, to improve performance on the specified task. We publish the code for deriving control vectors and analyzing model representations. The method allows us to improve performance on reasoning benchmarks and assess how control vectors influence the final logit distribution of a model via metrics such as KL divergence and entropy. We apply control vectors to Mistral-7B-Instruct and a range of Pythia models on an inductive, a deductive and mathematical reasoning task. We show that an LLM can, to a certain degree, be controlled to improve its perceived reasoning ability by modulating activations. The intervention is dependent upon the ability to reliably extract the model's typical state when correctly solving a task. Our results suggest that reasoning performance can be modulated in the same manner as other information-processing tasks performed by LLMs and demonstrate that we are capable of improving performance on specific tasks via a simple intervention on the residual stream with no additional training.

1182. Interpreting Language Reward Models via Contrastive Explanations

链接：https://iclr.cc/virtual/2025/poster/28716 abstract： Reward models (RMs) are a crucial component in the alignment of large language models' (LLMs) outputs with human values. RMs approximate human preferences over possible LLM responses to the same prompt by predicting and comparing reward scores. However, as they are typically modified versions of LLMs with scalar output heads, RMs are large black boxes whose predictions are not explainable. More transparent RMs would enable improved trust in the alignment of LLMs. In this work, we propose to use contrastive explanations to explain any binary response comparison made by an RM. Specifically, we generate a diverse set of new comparisons similar to the original one to characterise the RM's local behaviour. The perturbed responses forming the new comparisons are generated to explicitly modify manually specified high-level evaluation attributes, on which analyses of RM behaviour are grounded. In quantitative experiments, we validate the effectiveness of our method for finding high-quality contrastive explanations. We then showcase the qualitative usefulness of our method for investigating global sensitivity of RMs to each evaluation attribute, and demonstrate how representative examples can be automatically extracted to explain and compare behaviours of different RMs. We see our method as a flexible framework for RM explanation, providing a basis for more interpretable and trustworthy LLM alignment.

1183. Neural Context Flows for Meta-Learning of Dynamical Systems

链接：https://iclr.cc/virtual/2025/poster/30735 abstract： Neural Ordinary Differential Equations (NODEs) often struggle to adapt to new dynamic behaviors caused by parameter changes in the underlying physical system, even when these dynamics are similar to previously observed behaviors. This problem becomes more challenging when the changing parameters are unobserved, meaning their value or influence cannot be directly measured when collecting data. To address this issue, we introduce Neural Context Flow (NCF), a robust and interpretable Meta-Learning framework that includes uncertainty estimation. NCF uses Taylor expansion to enable contextual self-modulation, allowing context vectors to influence dynamics from other domains while also modulating themselves. After establishing theoretical guarantees, we empirically test NCF and compare it to related adaptation methods. Our results show that NCF achieves state-of-the-art Out-of-Distribution performance on 5 out of 6 linear and non-linear benchmark problems. Through extensive experiments, we explore the flexible model architecture of NCF and the encoded representations within the learned context vectors. Our findings highlight the potential implications of NCF for foundational models in the physical sciences, offering a promising approach to improving the adaptability and generalization of NODEs in various scientific applications.

1184. From Complexity to Clarity: Analytical Expressions of Deep Neural Network Weights via Clifford Algebra and Convexity

链接：https://iclr.cc/virtual/2025/poster/31470 abstract： In this paper, we introduce a novel analysis of neural networks based on geometric (Clifford) algebra and convex optimization. We show that optimal weights of deep ReLU neural networks are given by the wedge product of training samples when trained with standard regularized loss. Furthermore, the training problem reduces to convex optimization over wedge product features, which encode the geometric structure of the training dataset. This structure is given in terms of signed volumes of triangles and parallelotopes generated by data vectors. The convex problem finds a small subset of samples via ℓ_1 regularization to discover only relevant wedge product features. Our analysis provides a novel perspective on the inner workings of deep neural networks and sheds light on the role of the hidden layers.

1185. PnP-Flow: Plug-and-Play Image Restoration with Flow Matching

链接：https://iclr.cc/virtual/2025/poster/30965 abstract： In this paper, we introduce Plug-and-Play (PnP) Flow Matching, an algorithm for solving imaging inverse problems. PnP methods leverage the strength of pre-trained denoisers, often deep neural networks, by integrating them in optimization schemes. While they achieve state-of-the-art performance on various inverse

problems in imaging, PnP approaches face inherent limitations on more generative tasks like inpainting. On the other hand, generative models such as Flow Matching pushed the boundary in image sampling yet lack a clear method for efficient use in image restoration. We propose to combine the PnP framework with Flow Matching (FM) by defining a time-dependent denoiser using a pre-trained FM model. Our algorithm alternates between gradient descent steps on the data-fidelity term, reprojections onto the learned FM path, and denoising. Notably, our method is computationally efficient and memory-friendly, as it avoids backpropagation through ODEs and trace computations. We evaluate its performance on denoising, super-resolution, deblurring, and inpainting tasks, demonstrating superior results compared to existing PnP algorithms and Flow Matching based state-of-the-art methods. Code available at https://github.com/annegnx/PnP-Flow.

1186. Minimal Variance Model Aggregation: A principled, non-intrusive, and versatile integration of black box models

链接：https://iclr.cc/virtual/2025/poster/28797 abstract： Whether deterministic or stochastic, models can be viewed as functions designed to approximate a specific quantity of interest. We introduce Minimal Empirical Variance Aggregation (MEVA), a data-driven framework that integrates predictions from various models, enhancing overall accuracy by leveraging the individual strengths of each. This non-intrusive, model-agnostic approach treats the contributing models as black boxes and accommodates outputs from diverse methodologies, including machine learning algorithms and traditional numerical solvers.We advocate for a point-wise linear aggregation process and consider two methods for optimizing this aggregate: Minimal Error Aggregation (MEA), which minimizes the prediction error, and Minimal Variance Aggregation (MVA), which focuses on reducing variance. We prove a theorem showing that MVA can be more robustly estimated from data than MEA, making MEVA superior to Minimal Empirical Error Aggregation (MEEA). Unlike MEEA, which interpolates target values directly, MEVA formulates aggregation as an error estimation problem, which can be performedusing any backbone learning paradigm. We demonstrate the versatility and effectiveness of our framework across various applications, including data science and partial differential equations, illustrating its ability to significantly enhance both robustness and accuracy.

1187. Scaling and evaluating sparse autoencoders

链接：https://iclr.cc/virtual/2025/poster/28040 abstract： Sparse autoencoders provide a promising unsupervised approach for extracting interpretable features from a language model by reconstructing activations from a sparse bottleneck layer. Since language models learn many concepts, autoencoders need to be very large to recover all relevant features. However, studying the properties of autoencoder scaling is difficult due to the need to balance reconstruction and sparsity objectives and the presence of dead latents. We propose using k-sparse autoencoders [Makhzani and Frey, 2013] to directly control sparsity, simplifying tuning and improving the reconstruction-sparsity frontier. Additionally, we find modifications that result in few dead latents, even at the largest scales we tried. Using these techniques, we find clean scaling laws with respect to autoencoder size and sparsity. We also introduce several new metrics for evaluating feature quality based on the recovery of hypothesized features, the explainability of activation patterns, and the sparsity of downstream effects. These metrics all generally improve with autoencoder size. To demonstrate the scalability of our approach, we train a 16 million latent autoencoder on GPT-4 activations for 40 billion tokens. We release training code and autoencoders for open-source models, as well as a visualizer.

1188. MIRACLE 3D: Memory-efficient Integrated Robust Approach for Continual Learning on 3D Point Clouds via Shape Model Construction

链接：https://iclr.cc/virtual/2025/poster/30637 abstract： In this paper, we introduce a novel framework for memory-efficient and privacy-preserving continual learning in 3D object classification. Unlike conventional memory-based approaches in continual learning that require storing numerous exemplars, our method constructs a compact shape model for each class, retaining only the mean shape along with a few key modes of variation. This strategy not only enables the generation of diverse training samples while drastically reducing memory usage but also enhances privacy by eliminating the need to store original data. To further improve model robustness against input variations—an issue common in 3D domains due to the absence of strong backbones and limited training data—we incorporate Gradient Mode Regularization. This technique enhances model stability and broadens classification margins, resulting in accuracy improvements. We validate our approach through extensive experiments on the ModelNet40, ShapeNet, and ScanNet datasets, where we achieve state-of-the-art performance. Notably, our method consumes only 15% of the memory required by competing methods on the ModelNet40 and ShapeNet, while achieving comparable performance on the challenging ScanNet dataset with just 8.5% of the memory. These results underscore the scalability, effectiveness, and privacy-preserving strengths of our framework for 3D object classification.

1189. Distributed Speculative Inference (DSI): Speculation Parallelism for Provably Faster Lossless Language Model Inference

链接：https://iclr.cc/virtual/2025/poster/29058 abstract： This paper introduces distributed speculative inference (DSI), a novel inference algorithm that is provably faster than speculative inference (SI) [leviathan2023, chen2023, miao2024, sun2025, timor2025] and standard autoregressive inference (non-SI). Like other SI algorithms, DSI operates on frozen language models (LMs), requiring no training or architectural modifications, and it preserves the target distribution. Prior studies on SI have demonstrated empirical speedups over non-SI—but rely on sufficiently fast and accurate drafters, which are often unavailable in practice. We identify a gap where SI can be slower than non-SI if drafters are too slow or inaccurate. We close this gap by

proving that DSI is faster than both SI and non-SI—given any drafters. DSI is therefore not only faster than SI, but also unlocks the acceleration of LMs for which SI fails. DSI leverages speculation parallelism (SP), a novel type of task parallelism, to orchestrate target and drafter instances that overlap in time, establishing a new foundational tradeoff between computational resources and latency. Our simulations show that DSI is 1.29-1.92x faster than SI in single-node setups for various off-the-shelf LMs and tasks. We open-source all our code.

1190. Posterior-Mean Rectified Flow: Towards Minimum MSE Photo-Realistic Image Restoration

链接：https://iclr.cc/virtual/2025/poster/28765 abstract： Photo-realistic image restoration algorithms are typically evaluated by distortion measures (e.g., PSNR, SSIM) and by perceptual quality measures (e.g., FID, NIQE), where the desire is to attain the lowest possible distortion without compromising on perceptual quality. To achieve this goal, current methods commonly attempt to sample from the posterior distribution, or to optimize a weighted sum of a distortion loss (e.g., MSE) and a perceptual quality loss (e.g., GAN). Unlike previous works, this paper is concerned specifically with the optimal estimator that minimizes the MSE under a constraint of perfect perceptual index, namely where the distribution of the reconstructed images is equal to that of the ground-truth ones. A recent theoretical result shows that such an estimator can be constructed by optimally transporting the posterior mean prediction (MMSE estimate) to the distribution of the ground-truth images. Inspired by this result, we introduce Posterior-Mean Rectified Flow (PMRF), a simple yet highly effective algorithm that approximates this optimal estimator. In particular, PMRF first predicts the posterior mean, and then transports the result to a high-quality image using a rectified flow model that approximates the desired optimal transport map. We investigate the theoretical utility of PMRF and demonstrate that it consistently outperforms previous methods on a variety of image restoration tasks.

1191. Controllable Generation via Locally Constrained Resampling

链接：https://iclr.cc/virtual/2025/poster/30751 abstract： Autoregressive models have demonstrated an unprecedented ability at modeling the intricaciesof natural language. However, they continue to struggle with generating complex outputs that adhere tological constraints. Sampling from a fully-independent distribution subject to a constraint is hard. Sampling from an autoregressive distribution subject to a constraint is doubly hard: We have to contend not only with the hardness of the constraint but also the distribution's lack of structure. We propose a tractable probabilistic approach that performs Bayesian conditioning to draw samples subject to a constraint. By factoring in information about the entire sequence, our approach offers better contextual awareness during constrained generation compared to current greedy approaches. Starting from a model sample, we induce a local, factorized distribution which we cantractably condition on the constraint. To generate samples that satisfy the constraint, we sample from the conditional distribution,correct for biases in the sample weights, and resample. The resulting samples closely approximate the target distribution and are guaranteed to satisfy the constraints. We evaluate our approach on several tasks, including LLM detoxification and solving Sudoku puzzles. We show that by disallowing a list of toxic expressions our approach is able to steer the model's outputs away from toxic generations, outperforming similar approaches to detoxification. We also show that our approach achieves a perfect accuracy on Sudoku, compared to less than 50% for GPT4-o and Gemini 1.5.

1192. Articulate-Anything: Automatic Modeling of Articulated Objects via a Vision-Language Foundation Model

链接：https://iclr.cc/virtual/2025/poster/28149 abstract： Interactive 3D simulated objects are crucial in AR/VR, animations, and robotics, driving immersive experiences and advanced automation.However, creating these articulated objects requires extensive human effort and expertise, limiting their broader applications. To overcome this challenge, we present Articulate-Anything, a system that automates the articulation of diverse, complex objects from many input modalities, including text, images, and videos. Articulate-Anything leverages vision-language models (VLMs) to generate code that can be compiled into an interactable digital twin for use in standard 3D simulators. Our system exploits existing 3D asset datasets via a mesh retrieval mechanism, along with an actor-critic system that iteratively proposes, evaluates, and refines solutions for articulating the objects, self-correcting errors to achieve a robust out- come. Qualitative evaluations demonstrate Articulate-Anything's capability to articulate complex and even ambiguous object affordances by leveraging rich grounded inputs. In extensive quantitative experiments on the standard PartNet-Mobility dataset, Articulate-Anything substantially outperforms prior work, increasing the success rate from 8.7-11.6\% to 75\% and setting a new bar for state-of-art performance. We further showcase the utility of our generated assets by using them to train robotic policies for fine-grained manipulation tasks that go beyond basic pick and place.

1193. Deep Networks Learn Features From Local Discontinuities in the Label Function

链接：https://iclr.cc/virtual/2025/poster/30974 abstract： Deep neural networks outperform kernel machines on several datasets due to feature learning that happens during gradient descent training. In this paper, we analyze the mechanism through which feature learning happens and use a notion of features that corresponds to discontinuities in the true label function. We hypothesize that the core feature learning mechanism is label function discontinuities attracting model function discontinuities during training. To test this hypothesis, we perform experiments on classification data where the true label function is given by an oblique decision tree. This setup allows easy enumeration of label function discontinuities, while still remaining intractable for

static kernel/linear methods. We then design/construct a novel deep architecture called a Deep Linearly Gated Network (DLGN), whose discontinuities in the input space can be easily enumerated. In this setup, we provide supporting evidence demonstrating the movement of model function discontinuities towards the label function discontinuities during training. The easy enumerability of discontinuities in the DLGN also enables greater mechanistic interpretability. We demonstrate this by extracting the parameters of a high-accuracy decision tree from the parameters of a DLGN. We also show that the DLGN is competitive with ReLU networks and other tree-learning algorithms on several real-world tabular datasets.

1194. Deep Compression Autoencoder for Efficient High-Resolution Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/27858 abstract： We present Deep Compression Autoencoder (DC-AE), a new family of autoencoders for accelerating high-resolution diffusion models. Existing autoencodes have demonstrated impressive results at a moderate spatial compression ratio (e.g., 8x), but fail to maintain satisfactory reconstruction accuracy for high spatial compression ratios (e.g., 64x). We address this challenge by introducing two key techniques: (1) Residual Autoencoding, where we design our models to learn residuals based on the space-to-channel transformed features to alleviate the optimization difficulty of high spatial-compression autoencoders; (2) Decoupled High-Resolution Adaptation, an efficient decoupled three-phase training strategy for mitigating the generalization penalty of high spatial-compression autoencoders. With these designs, we improve the autoencoder's spatial compression ratio up to 128 while maintaining the reconstruction quality. Applying our DC-AE to latent diffusion models, we achieve significant speedup without accuracy drop. For example, on ImageNet 512x512, our DC-AE provides 19.1x inference speedup and 17.9x training speedup on H100 GPU for UViT-H while achieving a better FID, compared with the widely used SD-VAE-f8 autoencoder.

1195. Linear Mode Connectivity in Differentiable Tree Ensembles

链接：https://iclr.cc/virtual/2025/poster/29452 abstract： Linear Mode Connectivity (LMC) refers to the phenomenon that performance remains consistent for linearly interpolated models in the parameter space. For independently optimized model pairs from different random initializations, achieving LMC is considered crucial for understanding the stable success of the non-convex optimization in modern machine learning models and for facilitating practical parameter-based operations such as model merging. While LMC has been achieved for neural networks by considering the permutation invariance of neurons in each hidden layer, its attainment for other models remains an open question. In this paper, we first achieve LMC for soft tree ensembles, which are tree-based differentiable models extensively used in practice. We show the necessity of incorporating two invariances: subtree flip invariance and splitting order invariance, which do not exist in neural networks but are inherent to tree architectures, in addition to permutation invariance of trees. Moreover, we demonstrate that it is even possible to exclude such additional invariances while keeping LMC by designing decision list-based tree architectures, where such invariances do not exist by definition. Our findings indicate the significance of accounting for architecture-specific invariances in achieving LMC.

1196. Text4Seg: Reimagining Image Segmentation as Text Generation

链接：https://iclr.cc/virtual/2025/poster/27892 abstract： Multimodal Large Language Models (MLLMs) have shown exceptional capabilities in vision-language tasks; however, effectively integrating image segmentation into these models remains a significant challenge. In this paper, we introduce Text4Seg, a novel text-as-mask paradigm that casts image segmentation as a text generation problem, eliminating the need for additional decoders and significantly simplifying the segmentation process. Our key innovation is semantic descriptors, a new textual representation of segmentation masks where each image patch is mapped to its corresponding text label. This unified representation allows seamless integration into the auto-regressive training pipeline of MLLMs for easier optimization. We demonstrate that representing an image with 16×16 semantic descriptors yields competitive segmentation performance. To enhance efficiency, we introduce the Row-wise Run-Length Encoding (R-RLE), which compresses redundant text sequences, reducing the length of semantic descriptors by 74\% and accelerating inference by $3\times$, without compromising performance. Extensive experiments across various vision tasks, such as referring expression segmentation and comprehension, show that Text4Seg achieves state-of-the-art performance on multiple datasets by fine-tuning different MLLM backbones. Our approach provides an efficient, scalable solution for vision-centric tasks within the MLLM framework.

1197. SIMPL: Scalable and hassle-free optimisation of neural representations from behaviour

链接：https://iclr.cc/virtual/2025/poster/30670 abstract： Neural activity in the brain is known to encode low-dimensional, time-evolving, behaviour-related variables. A long-standing goal of neural data analysis has been to identify these variables and their mapping to neural activity. A productive and canonical approach has been to simply visualise neural "tuning curves" as a function of behaviour. However, significant discrepancies between behaviour and the true latent variables -- such as an agent thinking of position Y whilst located at position X -- distort and blur the tuning curves, decreasing their interpretability. To address this, latent variable models propose to learn the latent variable from data; these are typically expensive, hard to tune, or scale poorly, complicating their adoption. Here we propose SIMPL (Scalable Iterative Maximization of Population-coded Latents), an EM-style algorithm which iteratively optimises latent variables and tuning curves. SIMPL is fast, scalable and exploits behaviour as an initial condition to further improve convergence and identifiability. It can accurately recover latent variables in spatial and non-spatial tasks. When applied to a large hippocampal dataset SIMPL converges on smaller, more numerous, and more uniformly

sized place fields than those based on behaviour, suggesting the brain may encode space with greater resolution than previously thought.

1198. Building Blocks of Differentially Private Training

链接：https://iclr.cc/virtual/2025/poster/31339 abstract： In this blog, we introduce the building blocks of training a neural network in a differentially private way.

1199. Language Models Trained to do Arithmetic Predict Human Risky and Intertemporal Choice

链接：https://iclr.cc/virtual/2025/poster/29516 abstract： The observed similarities in the behavior of humans and Large Language Models (LLMs) have prompted researchers to consider the potential of using LLMs as models of human cognition. However, several significant challenges must be addressed before LLMs can be legitimately regarded as cognitive models. For instance, LLMs are trained on far more data than humans typically encounter, and may have been directly trained on human data in specific cognitive tasks or aligned with human preferences. Consequently, the origins of these behavioral similarities are not well understood. In this paper, we propose a novel way to enhance the utility of language models as cognitive models. This approach involves (i) leveraging computationally equivalent tasks that both a language model and a rational agent need to master for solving a cognitive problem and (ii) examining the specific task distributions required for a language model to exhibit human-like behaviors. We apply this approach to decision-making -- specifically risky and intertemporal choice -- where the key computationally equivalent task is the arithmetic of expected value calculations. We show that a small language model pretrained on an ecologically valid arithmetic dataset, which we call Arithmetic-GPT, predicts human behavior better than many traditional cognitive models. Pretraining language models on ecologically valid arithmetic datasets is sufficient to produce a strong correspondence between these models and human decision-making. Our results also suggest that language models used as cognitive models should be carefully investigated via ablation studies of the pretraining data.

1200. Formation of Representations in Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29863 abstract： Understanding neural representations will help open the black box of neural networks and advance our scientific understanding of modern AI systems. However, how complex, structured, and transferable representations emerge in modern neural networks has remained a mystery. Building on previous results, we propose the Canonical Representation Hypothesis (CRH), which posits a set of six alignment relations to universally govern the formation of representations in most hidden layers of a neural network. Under the CRH, the latent representations (R), weights (W), and neuron gradients (G) become mutually aligned during training. This alignment implies that neural networks naturally learn compact representations, where neurons and weights are invariant to task-irrelevant transformations. We then show that the breaking of CRH leads to the emergence of reciprocal power-law relations between R, W, and G, which we refer to as the Polynomial Alignment Hypothesis (PAH). We present a minimal-assumption theory proving that the balance between gradient noise and regularization is crucial for the emergence of the canonical representation. The CRH and PAH lead to an exciting possibility of unifying major key deep learning phenomena, including neural collapse and the neural feature ansatz, in a single framework.