3601. Provable Robust Overfitting Mitigation in Wasserstein Distributionally Robust Optimization

链接：https://iclr.cc/virtual/2025/poster/28096 abstract： Wasserstein distributionally robust optimization (WDRO) optimizes against worst-case distributional shifts within a specified uncertainty set, leading to enhanced generalization on unseen adversarial examples, compared to standard adversarial training which focuses on pointwise adversarial perturbations. However, WDRO still suffers fundamentally from the robust overfitting problem, as it does not consider statistical error. We address this gap by proposing a novel robust optimization framework under a new uncertainty set for adversarial noise via Wasserstein distance and statistical error via Kullback-Leibler divergence, called the Statistically Robust WDRO. We establish a robust generalization bound for the new optimization framework, implying that out-of-distribution adversarial performance is at least as good as the statistically robust training loss with high probability. Furthermore, we derive conditions under which Stackelberg and Nash equilibria exist between the learner and the adversary, giving an optimal robust model in certain sense.Finally, through extensive experiments, we demonstrate that our method significantly mitigates robust overfitting and enhances robustness within the framework of WDRO.

3602. Complementary Label Learning with Positive Label Guessing and Negative Label Enhancement

链接：https://iclr.cc/virtual/2025/poster/29999 abstract： Complementary label learning (CLL) is a weakly supervised learning paradigm that constructs a multi-class classifier only with complementary labels, specifying classes that the instance does not belong to. We reformulate CLL as an inverse problem that infers the full label information from the output space information. To be specific, we propose to split the inverse problem into two subtasks: positive label guessing (PLG) and negative label enhancement (NLE), collectively called PLNL. Specifically, we use well-designed criteria for evaluating the confidence of the model output, accordingly divide the training instances into three categories: highly-confident, moderately-confident and under-confident. For highly-confident instances, we perform PLG to assign them pseudo labels for supervised training. For moderately-confident and under-confident instances, we perform NLE by enhancing their complementary label set at different levels and train them with the augmented complementary labels iteratively. In addition, we unify PLG and NLE into a consistent framework, in which we can view all the pseudo-labeling-based methods from the perspective of negative label recovery. We prove that the error rates of both PLG and NLE are upper bounded, and based on that we can construct a classifier consistent with that learned by clean full labels. Extensive experiments demonstrate the superiority of PLNL over the state-of-the-art CLL methods, e.g., on STL-10, we increase the classification accuracy from 34.96\% to 55.25\%. The source code is available at https://github.com/yhli-ml/PLNL.

3603. InstaRevive: One-Step Image Enhancement via Dynamic Score Matching

链接：https://iclr.cc/virtual/2025/poster/30310 abstract： Image enhancement finds wide-ranging applications in real-world scenarios due to complex environments and the inherent limitations of imaging devices. Recent diffusion-based methods yield promising outcomes but necessitate prolonged and computationally intensive iterative sampling. In response, we propose InstaRevive, a straightforward yet powerful image enhancement framework that employs score-based diffusion distillation to harness potent generative capability and minimize the sampling steps. To fully exploit the potential of the pre-trained diffusion model, we devise a practical and effective diffusion distillation pipeline using dynamic noise control to address inaccuracies in updating direction during score matching. Our noise control strategy enables a dynamic diffusing scope, facilitating precise learning of denoising trajectories within the diffusion model and ensuring accurate distribution matching gradients during training. Additionally, to enrich guidance for the generative power, we incorporate textual prompts via image captioning as auxiliary conditions, fostering further exploration of the diffusion model. Extensive experiments substantiate the efficacy of our framework across a diverse array of challenging tasks and datasets, unveiling the compelling efficacy and efficiency of InstaRevive in delivering high-quality and visually appealing results.

3604. Physics of Language Models: Part 3.2, Knowledge Manipulation

链接：https://iclr.cc/virtual/2025/poster/28369 abstract： Language models can store vast factual knowledge, yet their ability to flexibly use this knowledge for downstream tasks (e.g., via instruction finetuning) remains questionable. This paper investigates four fundamental knowledge manipulation tasks: \textbf{retrieval} (e.g., "What is person A's attribute X?"), \textbf{classification} (e.g., "Is A's attribute X even or odd?"), \textbf{comparison} (e.g., "Is A greater than B in attribute X?"), and \textbf{inverse search} (e.g., "Which person's attribute X equals T?").We show that language models excel in knowledge retrieval but struggle even in the simplest classification or comparison tasks unless Chain of Thoughts (CoTs) are employed during both training and inference. Moreover, their performance in inverse knowledge search is virtually 0\%, regardless of the prompts.Our primary contribution is a \emph{controlled, synthetic experiment} that confirms these weaknesses are \emph{inherent} to language models: they cannot efficiently manipulate knowledge from pre-training data, even when such knowledge is perfectly stored in the models, despite adequate training and sufficient model size. Our findings also apply to modern pretrained language models such as GPT-4, thus giving rise to many Turing tests to distinguish Humans from contemporary AIs.

3605. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling

Laws

链接：https://iclr.cc/virtual/2025/poster/30313 abstract： Scaling laws describe the relationship between the size of language models and their capabilities. Unlike prior studies that evaluate a model's capability via loss or benchmarks, we estimate information-theoretically the number of knowledge \emph{bits} a model stores. We focus on factual knowledge represented as tuples, such as (USA, capital, Washington D.C.) from a Wikipedia page. Through multiple controlled datasets, we establish that language models can and only can store \emph{2 bits of knowledge per parameter, even when quantized to int8}, and such knowledge can be flexibly extracted for downstream applications. More broadly, we present 12 results on how (1) training duration, (2) model architecture, (3) quantization, (4) sparsity constraints such as MoE, and (5) data signal-to-noise ratio affect a model's knowledge storage capacity.

3606. Unhackable Temporal Reward for Scalable Video MLLMs

链接：https://iclr.cc/virtual/2025/poster/30267 abstract： In the pursuit of superior video-processing MLLMs, we have encountered a perplexing paradox: the “anti-scaling law”, where more data and larger models lead to worse performance. This study unmasks the culprit: “temporal hacking”, a phenomenon where models shortcut by fixating on select frames, missing the full video narrative. In this work, we systematically establish a comprehensive theory of temporal hacking, defining it from a reinforcement learning perspective, introducing the Temporal Perplexity (TPL) score to assess this misalignment, and proposing the Unhackable Temporal Rewarding (UTR) framework to mitigate the temporal hacking. Both theoretically and empirically, TPL proves to be a reliable indicator of temporal modeling quality, correlating strongly with frame activation patterns. Extensive experiments reveal that UTR not only counters temporal hacking but significantly elevates video comprehension capabilities. This work not only advances video-AI systems but also illuminates the critical importance of aligning proxy rewards with true objectives in MLLM development.

3607. GROOT-2: Weakly Supervised Multimodal Instruction Following Agents

链接：https://iclr.cc/virtual/2025/poster/29624 abstract： Developing agents that can follow multimodal instructions remains a fundamental challenge in robotics and AI. Although large-scale pre-training on unlabeled datasets has enabled agents to learn diverse behaviors, these agents often struggle with following instructions. While augmenting the dataset with instruction labels can mitigate this issue, acquiring such high-quality annotations at scale is impractical. To address this issue, we frame the problem as a semi-supervised learning task and introduce \agent, a multimodal instructable agent trained using a novel approach that combines weak supervision with latent variable models. Our method consists of two key components: constrained self-imitating, which utilizes large amounts of unlabeled demonstrations to enable the policy to learn diverse behaviors, and human intention alignment, which uses a smaller set of labeled demonstrations to ensure the latent space reflects human intentions. \agent’s effectiveness is validated across four diverse environments, ranging from video games to robotic manipulation, demonstrating its robust multimodal instruction-following capabilities.

3608. Knowledge Graph Finetuning Enhances Knowledge Manipulation in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28362 abstract： Despite the impressive performance of general large language models(LLMs), many of their applications in specific domains (e.g., low-data and knowledge-intensive) still confront significant challenges. Supervised fine-tuning (SFT)---where a general LLM is further trained on a small labeled dataset to adapt for specific tasks or domains---has shown great power for developing domain-specific LLMs. However, existing SFT data primarily consist of Question and Answer (Q&A) pairs, which poses a significant challenge for LLMs to comprehend the correlation and logic of knowledge underlying the Q&A. To address this challenge, we propose a conceptually flexible and general framework to boost SFT, namely Knowledge Graph-Driven Supervised Fine-Tuning (KG-SFT). The key idea of KG-SFT is to generate high-quality explanations for each Q&A pair via a structured knowledge graph to enhance the knowledge comprehension and manipulation of LLMs. Specifically, KG-SFT consists of three components: Extractor, Generator, and Detector. For a given Q&A pair, (i) Extractor first identifies entities within Q&A pairs and extracts relevant reasoning subgraphs from external KGs, (ii) Generator then produces corresponding fluent explanations utilizing these reasoning subgraphs, and (iii) finally, Detector performs sentence-level knowledge conflicts detection on these explanations to guarantee the reliability. KG-SFT focuses on generating high-quality explanations to improve the quality of Q&A pair, which reveals a promising direction for supplementing existing data augmentation methods. Extensive experiments on fifteen different domains and six different languages demonstrate the effectiveness of KG-SFT, leading to an accuracy improvement of up to 18% and an average of 8.7% in low-data scenarios.

3609. Computing Circuits Optimization via Model-Based Circuit Genetic Evolution

链接：https://iclr.cc/virtual/2025/poster/30053 abstract： Optimizing computing circuits such as multipliers and adders is a fundamental challenge in modern integrated circuit design. Recent efforts propose formulating this optimization problem as a reinforcement learning (RL) proxy task, offering a promising approach to search high-speed and area-efficient circuit design

solutions. However, we show that the RL-based formulation (proxy task) converges to a local optimal design solution (original task) due to the deceptive reward signals and incrementally localized actions in the RL-based formulation. To address this challenge, we propose a novel model-based circuit genetic evolution (MUTE) framework, which reformulates the problem as a genetic evolution process by proposing a grid-based genetic representation of design solutions. This novel formulation avoids misleading rewards by evaluating and improving generated solutions using the true objective value rather than proxy rewards. To promote globally diverse exploration, MUTE proposes a multi-granularity genetic crossover operator that recombines design substructures at varying column ranges between two grid-based genetic solutions. To the best of our knowledge, MUTE is the first to reformulate the problem as a circuit genetic evolution process, which enables effectively searching for global optimal design solutions. We evaluate MUTE on several fundamental computing circuits, including multipliers, adders, and multiply-accumulate circuits. Experiments on these circuits demonstrate that MUTE significantly Pareto-dominates state-of-the-art approaches in terms of both area and delay. Moreover, experiments demonstrate that circuits designed by MUTE well generalize to large-scale computation-intensive circuits as well.

3610. A Graph Enhanced Symbolic Discovery Framework For Efficient Logic Optimization

链接：https://iclr.cc/virtual/2025/poster/30414 abstract： The efficiency of Logic Optimization (LO) has become one of the key bottlenecks in chip design. To prompt efficient LO, previous studies propose using a key scoring function to predict and prune a large number of ineffective nodes of the LO heuristics. However, the existing scoring functions struggle to balance inference efficiency, interpretability, and generalization performance, which severely hinders their application to modern LO tools. To address this challenge, we propose a novel data-driven circuit symbolic learning framework, namely CMO, to learn lightweight, interpretable, and generalizable scoring functions. The major challenge of developing CMO is to discover symbolic functions that can well generalize to unseen circuits, i.e., the circuit symbolic generalization problem. Thus, the major technical contribution of CMO is the novel Graph Enhanced Symbolic Discovery framework, which distills dark knowledge from a well-designed Graph Neural Network (GNN) to enhance the generalization capability of the learned symbolic functions. To the best of our knowledge, CMO is the first graph-enhanced approach for discovering lightweight and interpretable symbolic functions that can well generalize to unseen circuits in LO. Experiments on three challenging circuit benchmarks show that the interpretable symbolic functions learned by CMO outperform previous state-of-the-art (SOTA) GPU-based and human-designed approaches in terms of inference efficiency and generalization capability. Moreover, we integrate CMO with the Mfs2 heuristic---one of the most time-consuming LO heuristics. The empirical results demonstrate that CMO significantly improves its efficiency while keeping comparable optimization performance when executed on a CPU-based machine, achieving up to 2.5× faster runtime.

3611. Beyond Content Relevance: Evaluating Instruction Following in Retrieval Models

链接：https://iclr.cc/virtual/2025/poster/29803 abstract： Instruction-following capabilities in large language models (LLMs) have progressed significantly, enabling more complex user interactions through detailed prompts. However, retrieval systems have not matched these advances, most of them still relies on traditional lexical and semantic matching techniques that fail to fully capture user intent. Recent efforts have introduced instruction-aware retrieval models, but these primarily focus on intrinsic content relevance, which neglects the importance of customized preferences for broader document-level attributes. This study evaluates the instruction-following capabilities of various retrieval models beyond content relevance, including LLM-based dense retrieval and reranking models. We develop InfoSearch, a novel retrieval evaluation benchmark spanning six document-level attributes: Audience, Keyword, Format, Language, Length, and Source, and introduce novel metrics -- Strict Instruction Compliance Ratio (SICR) and Weighted Instruction Sensitivity Evaluation (WISE) to accurately assess the models' responsiveness to instructions. Our findings indicate that although fine-tuning models on instruction-aware retrieval datasets and increasing model size enhance performance, most models still fall short of instruction compliance. We release our dataset and code on https://github.com/EIT-NLP/InfoSearch.

3612. Uni-Sign: Toward Unified Sign Language Understanding at Scale

链接：https://iclr.cc/virtual/2025/poster/31250 abstract： Sign language pre-training has gained increasing attention for its ability to enhance performance across various sign language understanding (SLU) tasks. However, existing methods often suffer from a gap between pre-training and fine-tuning, leading to suboptimal results. To address this, we propose Uni-Sign, a unified pre-training framework that eliminates the gap between pre-training and downstream SLU tasks through a large-scale generative pre-training strategy and a novel fine-tuning paradigm. First, we introduce CSL-News, a large-scale Chinese Sign Language (CSL) dataset containing 1,985 hours of video paired with textual annotations, which enables effective large-scale pre-training. Second, Uni-Sign unifies SLU tasks by treating downstream tasks as a single sign language translation (SLT) task during fine-tuning, ensuring seamless knowledge transfer between pre-training and fine-tuning. Furthermore, we incorporate a prior-guided fusion (PGF) module and a score-aware sampling strategy to efficiently fuse pose and RGB information, addressing keypoint inaccuracies and improving computational efficiency. Extensive experiments across multiple SLU benchmarks demonstrate that Uni-Sign achieves state-of-the-art performance across multiple downstream SLU tasks. Dataset and code are available at github.com/ZechengLi19/Uni-Sign.

3613. RMB: Comprehensively benchmarking reward models in LLM

alignment

链接：https://iclr.cc/virtual/2025/poster/28561 abstract： Reward models (RMs) guide the alignment of large language models (LLMs), steering them toward behaviors preferred by humans. Evaluating RMs is the key to better aligning LLMs. However, the current evaluation of RMs may not directly correspond to their alignment performance due to the limited distribution of evaluation data and evaluation methods that are not closely related to alignment objectives. To address these limitations, we propose RMB, a comprehensive RM benchmark that covers over 49 real-world scenarios and includes both pairwise and Best-of-N (BoN) evaluations to better reflect the effectiveness of RMs in guiding alignment optimization.We demonstrate a positive correlation between our benchmark and the downstream alignment task performance. Based on our benchmark, we conduct extensive analysis on the state-of-the-art RMs, revealing their generalization defects that were not discovered by previous benchmarks, and highlighting the potential of generative RMs. Furthermore, we delve into open questions in reward models, specifically examining the effectiveness of majority voting for the evaluation of reward models and analyzing the impact factors of generative RMs, including the influence of evaluation criteria and instructing methods. We will release our evaluation code and datasets upon publication.

3614. Proving Olympiad Inequalities by Synergizing LLMs and Symbolic Reasoning

链接：https://iclr.cc/virtual/2025/poster/30331 abstract： Large language models (LLMs) can prove mathematical theorems formally by generating proof steps (\textit{a.k.a.} tactics) within a proof system. However, the space of possible tactics is vast and complex, while the available training data for formal proofs is limited, posing a significant challenge to LLM-based tactic generation. To address this, we introduce a neuro-symbolic tactic generator that synergizes the mathematical intuition learned by LLMs with domain-specific insights encoded by symbolic methods. The key aspect of this integration is identifying which parts of mathematical reasoning are best suited to LLMs and which to symbolic methods. While the high-level idea of neuro-symbolic integration is broadly applicable to various mathematical problems, in this paper, we focus specifically on Olympiad inequalities (Figure~1). We analyze how humans solve these problems and distill the techniques into two types of tactics: (1) scaling, handled by symbolic methods, and (2) rewriting, handled by LLMs. In addition, we combine symbolic tools with LLMs to prune and rank the proof goals for efficient proof search. We evaluate our framework on 161 challenging inequalities from multiple mathematics competitions, achieving state-of-the-art performance and significantly outperforming existing LLM and symbolic approaches without requiring additional training data.

3615. Jailbreaking as a Reward Misspecification Problem

链接：https://iclr.cc/virtual/2025/poster/27994 abstract： The widespread adoption of large language models (LLMs) has raised concerns about their safety and reliability, particularly regarding their vulnerability to adversarial attacks. In this paper, we propose a new perspective that attributes this vulnerability to reward misspecification during the alignment process. This misspecification occurs when the reward function fails to accurately capture the intended behavior, leading to misaligned model outputs. We introduce a metric ReGap to quantify the extent of reward misspecification and demonstrate its effectiveness and robustness in detecting harmful backdoor prompts. Building upon these insights, we present ReMiss, a system for automated red teaming that generates adversarial prompts in a reward-misspecified space. ReMiss achieves state-of-the-art attack success rates on the AdvBench benchmark against various target aligned LLMs while preserving the human readability of the generated prompts. Furthermore, these attacks on open-source models demonstrate high transferability to closed-source models like GPT-4o and out-of-distribution tasks from HarmBench. Detailed analysis highlights the unique advantages of the proposed reward misspecification objective compared to previous methods, offering new insights for improving LLM safety and robustness.

3616. Vevo: Controllable Zero-Shot Voice Imitation with Self-Supervised Disentanglement

链接：https://iclr.cc/virtual/2025/poster/29148 abstract： The imitation of voice, targeted on specific speech attributes such as timbre and speaking style, is crucial in speech generation. However, existing methods rely heavily on annotated data, and struggle with effectively disentangling timbre and style, leading to challenges in achieving controllable generation, especially in zero-shot scenarios. To address these issues, we propose Vevo, a versatile zero-shot voice imitation framework with controllable timbre and style. Vevo operates in two core stages: (1) Content-Style Modeling: Given either text or speech's content tokens as input, we utilize an autoregressive transformer to generate the content-style tokens, which is prompted by a style reference; (2) Acoustic Modeling: Given the content-style tokens as input, we employ a flow-matching transformer to produce acoustic representations, which is prompted by a timbre reference. To obtain the content and content-style tokens of speech, we design a fully self-supervised approach that progressively decouples the timbre, style, and linguistic content of speech. Specifically, we adopt VQ-VAE as the tokenizer for the continuous hidden features of HuBERT. We treat the vocabulary size of the VQ-VAE codebook as the information bottleneck, and adjust it carefully to obtain the disentangled speech representations. Solely self-supervised trained on 60K hours of audiobook speech data, without any fine-tuning on style-specific corpora, Vevo matches or surpasses existing methods in accent and emotion conversion tasks. Additionally, Vevo’s effectiveness in zero-shot voice conversion and text-to-speech tasks further demonstrates its strong generalization and versatility. Audio samples are available at https://versavoice.github.io/.

3617. MagicPIG: LSH Sampling for Efficient LLM Generation

链接：https://iclr.cc/virtual/2025/poster/30638 abstract： Large language models (LLMs) with long context windows have gained significant attention. However, the KV cache, stored to avoid re-computation, becomes a bottleneck. Various dynamic sparse or TopK-based attention approximation methods have been proposed to leverage the common insight that attention is sparse. In this paper, we first show that TopK attention itself suffers from quality degradation in certain downstream tasks because attention is not always as sparse as expected. Rather than selecting the keys and values with the highest attention scores, sampling with theoretical guarantees can provide a better estimation for attention output. To make the sampling-based approximation practical in LLM generation, we propose MagicPIG, a heterogeneous system based on Locality Sensitive Hashing (LSH). MagicPIG significantly reduces the workload of attention computation while preserving high accuracy for diverse tasks. MagicPIG stores the LSH hash tables and runs the attention computation on the CPU, which allows it to serve longer contexts and larger batch sizes with high approximation accuracy. MagicPIG can improve decoding throughput by up to $5\times$ across various GPU hardware and achieve 54ms decoding latency on a single RTX 4090 for Llama-3.1-8B-Instruct model with a context of 96k tokens.

3618. D-FINE: Redefine Regression Task of DETRs as Fine-grained Distribution Refinement

链接：https://iclr.cc/virtual/2025/poster/29944 abstract： We introduce D-FINE, a powerful real-time object detector that achieves outstanding localization precision by redefining the bounding box regression task in DETR models. D-FINE comprises two key components: Fine-grained Distribution Refinement (FDR) and Global Optimal Localization Self-Distillation (GO-LSD). FDR transforms the regression process from predicting fixed coordinates to iteratively refining probability distributions, providing a fine-grained intermediate representation that significantly enhances localization accuracy. GO-LSD is a bidirectional optimization strategy that transfers localization knowledge from refined distributions to shallower layers through self-distillation, while also simplifying the residual prediction tasks for deeper layers. Additionally, D-FINE incorporates lightweight optimizations in computationally intensive modules and operations, achieving a better balance between speed and accuracy. Specifically, D-FINE-L / X achieves 54.0% / 55.8% AP on the COCO dataset at 124 / 78 FPS on an NVIDIA T4 GPU. When pretrained on Objects365, D-FINE-L / X attains 57.1% / 59.3% AP, surpassing all existing real-time detectors. Furthermore, our method significantly enhances the performance of a wide range of DETR models by up to 5.3% AP with negligible extra parameters and training costs. Our code and models: https://github.com/Peterande/D-FINE.

3619. Incremental Causal Effect for Time to Treatment Initialization

链接：https://iclr.cc/virtual/2025/poster/31239 abstract： We consider time to treatment initialization. This can commonly occur in preventive medicine, such as disease screening and vaccination; it can also occur with non-fatal health conditions such as HIV infection without the onset of AIDS. While traditional causal inference focused on ‘when to treat’ and its effects, including their possible dependence on subject characteristics, we consider the incremental causal effect when the intensity of time to treatment initialization is intervened upon. We provide identification of the incremental causal effect without the commonly required positivity assumption, as well as an estimation framework using inverse probability weighting. We illustrate our approach via simulation, and apply it to a rheumatoid arthritis study to evaluate the incremental effect of time to start methotrexate on joint pain.

3620. Benchmarking Multimodal Retrieval Augmented Generation with Dynamic VQA Dataset and Self-adaptive Planning Agent

链接：https://iclr.cc/virtual/2025/poster/29392 abstract： Multimodal Retrieval Augmented Generation (mRAG) plays an important role in mitigating the “hallucination” issue inherent in multimodal large language models (MLLMs). Although promising, existing heuristic mRAGs typically predefined fixed retrieval processes, which causes two issues: (1) Non-adaptive Retrieval Queries. (2) Overloaded Retrieval Queries. However, these flaws cannot be adequately reflected by current knowledge-seeking visual question answering (VQA) datasets, since the most required knowledge can be readily obtained with a standard two-step retrieval. To bridge the dataset gap, we first construct Dyn-VQA dataset, consisting of three types of ``dynamic'' questions, which require complex knowledge retrieval strategies variable in query, tool, and time: (1) Questions with rapidly changing answers. (2) Questions requiring multi-modal knowledge. (3) Multi-hop questions. Experiments on Dyn-VQA reveal that existing heuristic mRAGs struggle to provide sufficient and precisely relevant knowledge for dynamic questions due to their rigid retrieval processes. Hence, we further propose the first self-adaptive planning agent for multimodal retrieval, OmniSearch. The underlying idea is to emulate the human behavior in question solution which dynamically decomposes complex multimodal questions into sub-question chains with retrieval action. Extensive experiments prove the effectiveness of our OmniSearch, also provide direction for advancing mRAG. Code and dataset will be open-sourced.

3621. Graph Neural Ricci Flow: Evolving Feature from a Curvature Perspective

链接：https://iclr.cc/virtual/2025/poster/30813 abstract： Differential equations provide a dynamical perspective for understanding and designing graph neural networks (GNNs). By generalizing the discrete Ricci flow (DRF) to attributed graphs,

we can leverage a new paradigm for the evolution of node features with the help of curvature. We show that in the attributed graphs, DRF guarantees a vital property: The curvature of each edge concentrates toward zero over time. This property leads to two interesting consequences: 1) graph Dirichlet energy with bilateral bounds and 2) data-independent curvature decay rate. Based on these theoretical results, we propose the Graph Neural Ricci Flow (GNRF), a novel curvature-aware continuous-depth GNN. Compared to traditional curvature-based graph learning methods, GNRF is not limited to a specific curvature definition. It computes and adjusts time-varying curvature efficiently in linear time. We also empirically illustrate the operating mechanism of GNRF and verify that it performs excellently on diverse datasets.

3622. On Designing General and Expressive Quantum Graph Neural Networks with Applications to MILP Instance Representation

链接：https://iclr.cc/virtual/2025/poster/30170 abstract： Graph-structured data is ubiquitous, and graph learning models have recently been extended to address complex problems like mixed-integer linear programming (MILP). However, studies have shown that the vanilla message-passing based graph neural networks (GNNs) suffer inherent limitations in learning MILP instance representation, i.e., GNNs may map two different MILP instance graphs to the same representation. In this paper, we introduce an expressive quantum graph learning approach, leveraging quantum circuits to recognize patterns that are difficult for classical methods to learn. Specifically, the proposed General Quantum Graph Learning Architecture (GQGLA) is composed of a node feature layer, a graph message interaction layer, and an optional auxiliary layer. Its generality is reflected in effectively encoding features of nodes and edges while ensuring node permutation equivariance and flexibly creating different circuit structures for various expressive requirements and downstream tasks. GQGLA is well suited for learning complex graph tasks like MILP representation. Experimental results highlight the effectiveness of GQGLA in capturing and learning representations for MILPs. In comparison to traditional GNNs, GQGLA exhibits superior discriminative capabilities and demonstrates enhanced generalization across various problem instances, making it a promising solution for complex graph tasks.

3623. OmniBind: Large-scale Omni Multimodal Representation via Binding Spaces

链接：https://iclr.cc/virtual/2025/poster/28540 abstract： Recently, human-computer interaction with various modalities has shown promising applications, like GPT-4o and Gemini. Meanwhile, multimodal representation models have emerged as the foundation for these versatile multimodal understanding and generation pipeline. Models like CLIP, CLAP and ImageBind can map their specialized modalities into respective joint spaces. To construct a high-quality omni representation space that can be shared and expert in any modality, we propose to merge these advanced models into a unified space in scale. With this insight, we present \textbf{OmniBind}, advanced multimodal joint representation models via fusing knowledge of 14 pre-trained spaces, which support 3D, audio, image, video and language inputs. To alleviate the interference between different knowledge sources in integrated space, we dynamically assign weights to different spaces by learning routers with two objectives: cross-modal overall alignment and language representation decoupling. Notably, since binding and routing spaces only require lightweight networks, OmniBind is extremely training-efficient. Extensive experiments demonstrate the versatility and superiority of OmniBind as an omni representation model, highlighting its great potential for diverse applications, such as any-query and composable multimodal understanding.

3624. OmniSep: Unified Omni-Modality Sound Separation with Query-Mixup

链接：https://iclr.cc/virtual/2025/poster/30444 abstract： Query-based sound separation (QSS) effectively isolate sound signals that match the content of a given query, enhancing the understanding of audio data. However, most existing QSS methods rely on a single modality for separation, lacking the ability to fully leverage homologous but heterogeneous information across multiple modalities for the same sound signal. To address this limitation, we introduce Omni-modal Sound Separation (OmniSep), a novel framework capable of isolating clean soundtracks based on omni-modal queries, encompassing both single-modal and multi-modal composed queries. Specifically, we introduce the Query-Mixup strategy, which blends query features from different modalities during training. This enables OmniSep to optimize multiple modalities concurrently, effectively bringing all modalities under a unified framework for sound separation. We further enhance this flexibility by allowing queries to influence sound separation positively or negatively, facilitating the retention or removal of specific sounds as desired. Finally, OmniSep employs a retrieval-augmented approach known as Query-Aug, which enables open-vocabulary sound separation. Experimental evaluations on MUSIC, VGGSOUND-CLEAN+, and MUSIC-CLEAN+ datasets demonstrate effectiveness of OmniSep, achieving state-of-the-art performance in text-, image-, and audio-queried sound separation tasks. For samples and further information, please visit the demo page at \url{https://omnisep.github.io/}.

3625. On the Completeness of Invariant Geometric Deep Learning Models

链接：https://iclr.cc/virtual/2025/poster/30973 abstract： Invariant models, one important class of geometric deep learning models, are capable of generating meaningful geometric representations by leveraging informative geometric features in point clouds. These models are characterized by their simplicity, good experimental results and computational efficiency. However, their theoretical expressive power still remains unclear, restricting a deeper understanding of the potential of such models. In this work, we concentrate on characterizing the theoretical expressiveness of a wide range of invariant models under fully-connected conditions. We first rigorously characterize the expressiveness of the most classic invariant model, message-passing neural networks incorporating distance (DisGNN), restricting its unidentifiable cases to be only highly symmetric point clouds. We then

prove that GeoNGNN, the geometric counterpart of one of the simplest subgraph graph neural networks, can effectively break these corner cases' symmetry and thus achieve E(3)-completeness. By leveraging GeoNGNN as a theoretical tool, we further prove that: 1) most subgraph GNNs developed in traditional graph learning can be seamlessly extended to geometric scenarios with E(3)-completeness; 2) DimeNet, GemNet and SphereNet, three well-established invariant models, are also all capable of achieving E(3)-completeness. Our theoretical results fill the gap in the expressive power of invariant models, contributing to a rigorous and comprehensive understanding of their capabilities.

3626. MixEval-X: Any-to-any Evaluations from Real-world Data Mixture

链接：https://iclr.cc/virtual/2025/poster/28740 abstract： Perceiving and generating diverse modalities are crucial for AI models to effectively learn from and engage with real-world signals, necessitating reliable evaluations for their development. We identify two major issues in current evaluations: (1) inconsistent standards, shaped by different communities with varying protocols and maturity levels; and (2) significant query, grading, and generalization biases. To address these, we introduce MixEval-X, the first any-to-any, real-world benchmark designed to optimize and standardize evaluations across diverse input and output modalities. We propose multi-modal benchmark mixture and adaptation-rectification pipelines to reconstruct real-world task distributions, ensuring evaluations generalize effectively to real-world use cases. Extensive meta-evaluations show our approach effectively aligns benchmark samples with real-world task distributions. Meanwhile, MixEval-X's model rankings correlate strongly with that of crowd-sourced real-world evaluations (up to 0.98) while being much more efficient. We provide comprehensive leaderboards to rerank existing models and organizations and offer insights to enhance understanding of multi-modal evaluations and inform future research.

3627. RobustKV: Defending Large Language Models against Jailbreak Attacks via KV Eviction

链接：https://iclr.cc/virtual/2025/poster/30018 abstract： Jailbreak attacks circumvent LLMs' built-in safeguards by concealing harmful queries within adversarial prompts. While most existing defenses attempt to mitigate the effects of adversarial prompts, they often prove inadequate as adversarial prompts can take arbitrary, adaptive forms. This paper introduces RobustKV, a novel jailbreak defense that takes a fundamentally different approach by selectively removing critical tokens of harmful queries from key-value (KV) caches. Intuitively, for an adversarial prompt to be effective, its tokens must achieve sufficient `importance' (measured by attention scores), which consequently lowers the importance of tokens in the concealed harmful query. Therefore, by carefully evicting the KVs of low-ranked tokens, RobustKV minimizes the harmful query's presence in the KV cache, thus preventing the LLM from generating informative responses. Extensive evaluation using benchmark datasets and models demonstrates that RobustKV effectively counters state-of-the-art jailbreak attacks while maintaining the LLM's performance on benign queries. Notably, RobustKV creates an interesting effectiveness-evasiveness dilemma for the adversary, leading to its robustness against adaptive attacks.{(Warning: This paper contains potentially harmful content generated by LLMs.)}

3628. Entropy-based Activation Function Optimization: A Method on Searching Better Activation Functions

链接：https://iclr.cc/virtual/2025/poster/30823 abstract： The success of artificial neural networks (ANNs) hinges greatly on the judicious selection of an activation function, introducing non-linearity into network and enabling them to model sophisticated relationships in data. However, the search of activation functions has largely relied on empirical knowledge in the past, lacking theoretical guidance, which has hindered the identification of more effective activation functions. In this work, we offer a proper solution to such issue. Firstly, we theoretically demonstrate the existence of the worst activation function with boundary conditions (WAFBC) from the perspective of information entropy. Furthermore, inspired by the Taylor expansion form of information entropy functional, we propose the Entropy-based Activation Function Optimization (EAFO) methodology. EAFO methodology presents a novel perspective for designing static activation functions in deep neural networks and the potential of dynamically optimizing activation during iterative training. Utilizing EAFO methodology, we derive a novel activation function from ReLU, known as Correction Regularized ReLU (CRReLU). Experiments conducted with vision transformer and its variants on CIFAR-10, CIFAR-100 and ImageNet-1K datasets demonstrate the superiority of CRReLU over existing corrections of ReLU. Extensive empirical studies on task of large language model (LLM) fine-tuning, CRReLU exhibits superior performance compared to GELU, suggesting its broader potential for practical applications.

3629. BigDocs: An Open Dataset for Training Multimodal Models on Document and Code Tasks

链接：https://iclr.cc/virtual/2025/poster/29129 abstract： Multimodal AI has the potential to significantly enhance document-understanding tasks, such as processing receipts, understanding workflows, extracting data from documents, and summarizing reports. Code generation tasks that require long-structured outputs can also be enhanced by multimodality. Despite this, their use in commercial applications is often limited due to limited access to relevant training data and restrictive licensing, which hinders open access. To address these limitations, we introduce BigDocs-7.5M, a high-quality, open-access dataset comprising 7.5 million multimodal documents across 30 tasks. We use an efficient data curation process to ensure that our data is high quality and license-permissive. Our process emphasizes accountability, responsibility, and transparency through filtering rules, traceable metadata, and careful content analysis. Additionally, we introduce BigDocs-Bench,, a benchmark suite with 10

novel tasks where we carefully create datasets that reflect real-world use cases involving reasoning over Graphical User Interfaces (GUI) and code generation from images. Our experiments show that training with BigDocs-Bench, improves average performance up to 25.8% over closed-source GPT-4o in document reasoning and structured output tasks such as Screenshot2HTML or Image2Latex generation. Finally, human evaluations revealed that participants preferred the outputs from models trained with BigDocs over those from GPT-4o. This suggests that BigDocs can help both academics and the open-source community utilize and improve AI tools to enhance multimodal capabilities and document reasoning.

3630. From Isolated Conversations to Hierarchical Schemas: Dynamic Tree Memory Representation for LLMs

链接：https://iclr.cc/virtual/2025/poster/28447 abstract： Recent advancements in large language models have significantly improved their context windows, yet challenges in effective long-term memory management remain. We introduce MemTree, an algorithm that leverages a dynamic, tree-structured memory representation to optimize the organization, retrieval, and integration of information, akin to human cognitive schemas. MemTree organizes memory hierarchically, with each node encapsulating aggregated textual content, corresponding semantic embeddings, and varying abstraction levels across the tree's depths. Our algorithm dynamically adapts this memory structure by computing and comparing semantic embeddings of new and existing information to enrich the model’s context-awareness. This approach allows MemTree to handle complex reasoning and extended interactions more effectively than traditional memory augmentation methods, which often rely on flat lookup tables. Evaluations on benchmarks for multi-turn dialogue understanding and document question answering show that MemTree significantly enhances performance in scenarios that demand structured memory management.

3631. Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process

链接：https://iclr.cc/virtual/2025/poster/29518 abstract： Recent advances in language models have demonstrated their capability to solve mathematical reasoning problems, achieving near-perfect accuracy on grade-school level math benchmarks like GSM8K. In this paper, we formally study how language models solve these problems. We design a series of controlled experiments to address several fundamental questions: (1) Can language models truly develop reasoning skills, or do they simply memorize templates? (2) What is the model's hidden (mental) reasoning process? (3) Do models solve math questions using skills similar to or different from humans? (4) Do models trained on GSM8K-like datasets develop reasoning skills beyond those necessary for solving GSM8K problems? (5) What mental process causes models to make reasoning mistakes? (6) How large or deep must a model be to effectively solve GSM8K-level math questions?Our study uncovers many hidden mechanisms by which language models solve mathematical questions, providing insights that extend beyond current understandings of LLMs.

3632. Physics of Language Models: Part 2.2, How to Learn From Mistakes on Grade-School Math Problems

链接：https://iclr.cc/virtual/2025/poster/27648 abstract： Language models have demonstrated remarkable performance in solving reasoning tasks; however, even the strongest models still occasionally make reasoning mistakes. Recently, there has been active research aimed at improving reasoning accuracy, particularly by using pretrained language models to "self-correct" their mistakes via multi-round prompting. In this paper, we follow this line of work but focus on understanding the usefulness of incorporating ``error-correction'' data directly into the pretraining stage. This data consists of erroneous solution steps immediately followed by their corrections. Using a synthetic math dataset, we show promising results: this type of pretrain data can help language models achieve higher reasoning accuracy directly (i.e., through simple auto-regression, without multi-round prompting) compared to pretraining on the same amount of error-free data. We also delve into many details, such as (1) how this approach differs from beam search, (2) how such data can be prepared, (3) whether masking is needed on the erroneous tokens, (4) the amount of error required, (5) whether such data can be deferred to the fine-tuning stage, and many others.

3633. Retrieval Augmented Diffusion Model for Structure-informed Antibody Design and Optimization

链接：https://iclr.cc/virtual/2025/poster/29186 abstract： Antibodies are essential proteins responsible for immune responses in organisms, capable of specifically recognizing antigen molecules of pathogens. Recent advances in generative models have significantly enhanced rational antibody design. However, existing methods mainly create antibodies from scratch without template constraints, leading to model optimization challenges and unnatural sequences. To address these issues, we propose a retrieval-augmented diffusion framework, termed RADAb, for efficient antibody design. Our method leverages a set of structural homologous motifs that align with query structural constraints to guide the generative model in inversely optimizing antibodies according to desired design criteria. Specifically, we introduce a structure-informed retrieval mechanism that integrates these exemplar motifs with the input backbone through a novel dual-branch denoising module, utilizing both structural and evolutionary information. Additionally, we develop a conditional diffusion model that iteratively refines the optimization process by incorporating both global context and local evolutionary conditions. Our approach is agnostic to the choice of generative models. Empirical experiments demonstrate that our method achieves state-of-the-art performance in multiple antibody inverse folding and optimization tasks, offering a new perspective on biomolecular generative models.

3634. Humanizing the Machine: Proxy Attacks to Mislead LLM Detectors

链接：https://iclr.cc/virtual/2025/poster/29761 abstract： The advent of large language models (LLMs) has revolutionized the field of text generation, producing outputs that closely mimic human-like writing. Although academic and industrial institutions have developed detectors to prevent the malicious usage of LLM-generated texts, other research has doubt about the robustness of these systems. To stress test these detectors, we introduce a humanized proxy-attack (HUMPA) strategy that effortlessly compromises LLMs, causing them to produce outputs that align with human-written text and mislead detection systems. Our method attacks the source model by leveraging a reinforcement learning (RL) fine-tuned humanized small language model (SLM) in the decoding phase. Through an in-depth analysis, we demonstrate that our attack strategy is capable of generating responses that are indistinguishable to detectors, preventing them from differentiating between machine-generated and human-written text. We conduct systematic evaluations on extensive datasets using proxy-attacked open-source models, including Llama2-13B, Llama3-70B, and Mixtral-8x7B in both white- and black-box settings. Our findings show that the proxy-attack strategy effectively deceives the leading detectors, resulting in an average AUROC drop of 70.4% across multiple datasets, with a maximum drop of 95.0% on a single dataset. Furthermore, in cross-discipline scenarios, our strategy also bypasses these detectors, leading to a significant relative decrease of up to 90.9%, while in cross-language scenario, the drop reaches 91.3%. Despite our proxy-attack strategy successfully bypassing the detectors with such significant relative drops, we find that the generation quality of the attacked models remains preserved, even within a modest utility budget, when compared to the text produced by the original, unattacked source model.

3635. Montessori-Instruct: Generate Influential Training Data Tailored for Student Learning

链接：https://iclr.cc/virtual/2025/poster/30692 abstract： Synthetic data has been widely used to train large language models, but their generative nature inevitably introduces noisy, non-informative, and misleading learning signals. In this paper, we propose Montessori-Instruct, a novel data synthesis framework that tailors the data synthesis ability of the teacher language model toward the student language model's learning process. Specifically, we utilize local data influence of synthetic training data points on students to characterize students' learning preferences. Then, we train the teacher model with Direct Preference Optimization (DPO) to generate synthetic data tailored toward student learning preferences. Experiments with Llama3-8B-Instruct (teacher) and Llama3-8B (student) on Alpaca Eval and MT-Bench demonstrate that Montessori-Instruct significantly outperforms standard synthesis methods by 18.35\% and 46.24\% relatively. Our method also beats data synthesized by a stronger teacher model, GPT-4o. Further analysis confirms the benefits of teacher's learning to generate more influential training data in the student's improved learning, the advantages of local data influence in accurately measuring student preferences, and the robustness of Montessori-Instruct across different student models. Our code and data are open-sourced at https://github.com/cxcscmu/Montessori-Instruct.

3636. Tuning Timestep-Distilled Diffusion Model Using Pairwise Sample Optimization

链接：https://iclr.cc/virtual/2025/poster/28873 abstract： Recent advancements in timestep-distilled diffusion models have enabled high-quality image generation that rivals non-distilled multi-step models, but with significantly fewer inference steps. While such models are attractive for applications due to the low inference cost and latency, fine-tuning them with a naive diffusion objective would result in degraded and blurry outputs. An intuitive alternative is to repeat the diffusion distillation process with a fine-tuned teacher model, which produces good results but is cumbersome and computationally intensive: the distillation training usually requires magnitude higher of training compute compared to fine-tuning for specific image styles. In this paper, we present an algorithm named pairwise sample optimization (PSO), which enables the direct fine-tuning of an arbitrary timestep-distilled diffusion model. PSO introduces additional reference images sampled from the current time-step distilled model, and increases the relative likelihood margin between the training images and reference images. This enables the model to retain its few-step generation ability, while allowing for fine-tuning of its output distribution. We also demonstrate that PSO is a generalized formulation which be flexible extended to both offline-sampled and online-sampled pairwise data, covering various popular objectives for diffusion model preference optimization. We evaluate PSO in both preference optimization and other fine-tuning tasks, including style transfer and concept customization. We show that PSO can directly adapt distilled models to human-preferred generation with both offline and online-generated pairwise preference image data. PSO also demonstrates effectiveness in style transfer and concept customization by directly tuning timestep-distilled diffusion models.

3637. Episodic Memories Generation and Evaluation Benchmark for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/30855 abstract： Episodic memory -- the ability to recall specific events grounded in time and space -- is a cornerstone of human cognition, enabling not only coherent storytelling, but also planning and decision-making. Despite their remarkable capabilities, Large Language Models (LLMs) lack a robust mechanism for episodic memory: we argue that integrating episodic memory capabilities into LLM is essential for advancing AI towards human-like cognition, increasing their potential to reason consistently and ground their output in real-world episodic events, hence avoiding confabulations. To address this challenge, we introduce a comprehensive framework to model and evaluate LLM episodic memory capabilities. Drawing inspiration from cognitive science, we develop a structured approach to represent episodic

events, encapsulating temporal and spatial contexts, involved entities, and detailed descriptions. We synthesize a unique episodic memory benchmark, free from contamination, and release open source code and datasets to assess LLM performance across various recall and episodic reasoning tasks. Our evaluation of state-of-the-art models, including GPT-4 and Claude variants, Llama 3.1, and o1-mini, reveals that even the most advanced LLMs struggle with episodic memory tasks, particularly when dealing with multiple related events or complex spatio-temporal relationships -- even in contexts as short as 10k-100k tokens.

3638. B-STaR: Monitoring and Balancing Exploration and Exploitation in Self-Taught Reasoners

链接：https://iclr.cc/virtual/2025/poster/29776 abstract： In the absence of extensive human-annotated data for complex reasoning tasks, self-improvement -- where models are trained on their own outputs -- has emerged as a primary method for enhancing performance. Recently, the approach to self-improvement has shifted toward a more dynamic, online fashion through iterative training processes. However, the critical factors underlying the mechanism of these self-improving methods remain poorly understood, such as under what conditions self-improvement is effective, and what are the bottlenecks in the current iterations.In this work, we identify and propose methods to monitor two pivotal factors in this iterative process: (1) the model's ability to explore and generate high-quality responses among multiple candidates (exploration); and (2) the reliability of external rewards in selecting the best responses from the generated outputs (exploitation).These factors are inherently moving targets throughout the self-improvement cycles, yet their dynamics are rarely discussed in prior research -- It remains unclear what impedes continual model enhancement after only a few iterations. Using mathematical reasoning as a case study, we begin with a quantitative analysis to track the dynamics of exploration and exploitation, discovering that a model's exploratory capabilities rapidly deteriorate over iterations, and the effectiveness of exploiting external rewards diminishes as well due to shifts in distribution from the original policy.Motivated by these findings, we introduce B-STaR, a Self-Taught Reasoning framework that autonomously adjusts configurations across iterations to Balance exploration and exploitation, thereby optimizing the self-teaching effectiveness based on the current policy model and available rewards.Our experiments in mathematical reasoning demonstrate that B-STaR not only enhances the model's exploratory capabilities throughout training but also achieves a more effective balance between exploration and exploitation, leading to superior performance. Crucially, this work deconstructs the opaque nature of self-training algorithms, elucidating the interpretable dynamics throughout the process and highlighting current limitations for future research to address.

3639. Sort-free Gaussian Splatting via Weighted Sum Rendering

链接：https://iclr.cc/virtual/2025/poster/27749 abstract： Recently, 3D Gaussian Splatting (3DGS) has emerged as a significant advancement in 3D scene reconstruction, attracting considerable attention due to its ability to recover high-fidelity details while maintaining low complexity. Despite the promising results achieved by 3DGS, its rendering performance is constrained by its dependence on costly non-commutative alpha-blending operations. These operations mandate complex view dependent sorting operations that introduce computational overhead, especially on the resource-constrained platforms such as mobile phones. In this paper, we propose Weighted Sum Rendering, which approximates alpha blending with weighted sums, thereby removing the need for sorting. This simplifies implementation, delivers superior performance, and eliminates the ``popping'' artifacts caused by sorting. Experimental results show that optimizing a generalized Gaussian splatting formulation to the new differentiable rendering yields competitive image quality. The method was implemented and tested in a mobile device GPU, achieving on average $1.23\times$ faster rendering.

3640. Speculative Knowledge Distillation: Bridging the Teacher-Student Gap Through Interleaved Sampling

链接：https://iclr.cc/virtual/2025/poster/30392 abstract： Recent advances in knowledge distillation (KD) have enabled smaller student models to approach the performance of larger teacher models. However, popular methods such as supervised KD and on-policy KD, are adversely impacted by the knowledge gaps between teacher-student in practical scenarios. Supervised KD suffers from a distribution mismatch between training with a static dataset and inference over final student-generated outputs. Conversely, on-policy KD, which uses student-generated samples for training, can suffer from low-quality training examples with which teacher models are not familiar, resulting in inaccurate teacher feedback. To address these limitations, we introduce Speculative Knowledge Distillation (SKD), a novel approach that leverages cooperation between student and teacher models to generate high-quality training data on-the-fly while aligning with the student's inference-time distribution. In SKD, the student proposes tokens, and the teacher replaces poorly ranked ones based on its own distribution, transferring high-quality knowledge adaptively. We evaluate SKD on various text generation tasks, including translation, summarization, math, and instruction following, and show that SKD consistently outperforms existing KD methods across different domains, data sizes, and model initialization strategies.

3641. STAR: Stability-Inducing Weight Perturbation for Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30891 abstract： Humans can naturally learn new and varying tasks in a sequential manner. Continual learning is a class of learning algorithms that updates its learned model as it sees new data (on potentially new tasks) in a sequence. A key challenge in continual learning is that as the model is updated to learn new tasks, it becomes susceptible to \textit{catastrophic forgetting}, where knowledge of previously learned tasks is lost. A popular approach to

mitigate forgetting during continual learning is to maintain a small buffer of previously-seen samples, and to replay them during training. However, this approach is limited by the small buffer size and, while forgetting is reduced, it is still present. In this paper, we proposea novel loss function STAR that exploits the worst-case parameter perturbation that reduces the KL-divergence of model predictions with that of its local parameter neighborhood to promote stability and alleviate forgetting. STAR can be combined with almost any existing rehearsal-based methods as a plug-and-play component. We empirically show that STAR consistently improves performance of existing methods by up to $\sim15\%$ across varying baselines, and achieves superior or competitive accuracy to that of state-of-the-art methods aimed at improving rehearsal-based continual learning. Our implementation is available at https://github.com/Gnomy17/STAR_CL.

3642. OmnixR: Evaluating Omni-modality Language Models on Reasoning across Modalities

链接：https://iclr.cc/virtual/2025/poster/28621 abstract： We introduce \textbf{OmnixR}, an evaluation suite designed to benchmark state-of-the-art Omni-modality Language Models (OLMs), such as GPT-4o and Gemini. Evaluating OLMs, which integrate multiple modalities such as text, vision, and audio, presents unique challenges. Particularly, the user message might often consist of multiple modalities, such that OLMs have to establish holistic understanding and reasoning across modalities to accomplish the task.Existing benchmarks are limited to single-modality or dual-modality tasks (e.g., image+text or video+text), overlooking comprehensive multi-modal assessments of model reasoning.To address this, OmnixR offers two evaluation variants: (1) OmnixR-synth: a synthetic dataset generated automatically by translating text into multiple modalities—audio, images, video, and hybrids Omnify!. (2) OmnixR-real: a real-world dataset, manually curated and annotated by experts, for evaluating cross-modal reasoning in natural settings. OmnixR presents a unique evaluation towards assessing OLMs over a diverse mix of modalities, such as a question that involves video, audio, and text, providing a rigorous cross-modal reasoning testbed than any existing benchmarks.Our experiments find that all state-of-the-art OLMs struggles with OmnixR questions that require integrating information from multiple modalities to answer. Further analysis highlight differences in reasoning behavior and underscoring the challenges of omni-modal AI alignment.

3643. Self-Correcting Decoding with Generative Feedback for Mitigating Hallucinations in Large Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/28052 abstract： While recent Large Vision-Language Models (LVLMs) have shown remarkable performance in multi-modal tasks, they are prone to generating hallucinatory text responses that do not align with the given visual input, which restricts their practical applicability in real-world scenarios. In this work, inspired by the observation that the text-to-image generation process is the inverse of image-conditioned response generation in LVLMs, we explore the potential of leveraging text-to-image generative models to assist in mitigating hallucinations in LVLMs. We discover that generative models can offer valuable self-feedback for mitigating hallucinations at both the response and token levels. Building on this insight, we introduce self-correcting Decoding with Generative Feedback (DeGF), a novel training-free algorithm that incorporates feedback from text-to-image generative models into the decoding process to effectively mitigate hallucinations in LVLMs. Specifically, DeGF generates an image from the initial response produced by LVLMs, which acts as an auxiliary visual reference and provides self-feedback to verify and correct the initial response through complementary or contrastive decoding. Extensive experimental results validate the effectiveness of our approach in mitigating diverse types of hallucinations, consistently surpassing state-of-the-art methods across six benchmarks. Code is available at https://github.com/zhangce01/DeGF.

3644. OMG: Opacity Matters in Material Modeling with Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/28342 abstract： Decomposing geometry, materials and lighting from a set of images, namely inverse rendering, has been a long-standing problem in computer vision and graphics. Recent advances in neural rendering enable photo-realistic and plausible inverse rendering results. The emergence of 3D Gaussian Splatting has boosted it to the next level by showing real-time rendering potentials. An intuitive finding is that the models used for inverse rendering do not take into account the dependency of opacity w.r.t. material properties, namely cross section, as suggested by optics. Therefore, we develop a novel approach that adds this dependency to the modeling itself. Inspired by radiative transfer, we augment the opacity term by introducing a neural network that takes as input material properties to provide modeling of cross section and a physically correct activation function. The gradients for material properties are therefore not only from color but also from opacity, facilitating a constraint for their optimization. Therefore, the proposed method incorporates more accurate physical properties compared to previous works. We implement our method into 3 different baselines that use Gaussian Splatting for inverse rendering and achieve significant improvements universally in terms of novel view synthesis and material modeling.

3645. Preble: Efficient Distributed Prompt Scheduling for LLM Serving

链接：https://iclr.cc/virtual/2025/poster/28456 abstract： Prompts to large language models (LLMs) have evolved beyond simple user questions.For LLMs to solve complex problems, today’s practices are to include domain-specificinstructions, illustration of tool usages, and/or long context such as textbook chapters inprompts. As such, many parts of prompts are repetitive across requests. Recent workspropose to cache and reuse KV state of prompts. However, they are all confined to a single-GPU optimization, while production LLM serving systems are distributed by nature.This paper proposes Preble, the first

distributed LLM serving platform that targets and op-timizes for prompt sharing. We designed a distributed scheduling system that co-optimizesKV state reuse and computation load-balancing with a new scheduling algorithm and ahierarchical scheduling mechanism. Our evaluation of Preble with real workloads and re-quest arrival patterns on two open-source LLMs shows that Preble outperforms the SOTAserving systems by 1.5× to 14.5× on average latency and 2× to 10× on p99 latency.

3646. Learning Hierarchical Polynomials of Multiple Nonlinear Features

链接：https://iclr.cc/virtual/2025/poster/29467 abstract： In deep learning theory, a critical question is to understand how neural networks learn hierarchical features. In this work, we study the learning of hierarchical polynomials of multiple nonlinear features using three-layer neural networks. We examine a broad class of functions of the form $f^{\star}=g^{\star}\circ \mathbf{p}$, where $\mathbf{p}:\mathbb{R}^{d} \rightarrow \mathbb{R}^{r}$ represents multiple quadratic features with $r \ll d$ and $g^{\star}:\mathbb{R}^{r}\rightarrow \mathbb{R}$ is a polynomial of degree p. This can be viewed as a nonlinear generalization of the multi-index model, and also an expansion upon previous work on nonlinear feature learning that focused only on a single feature (i.e. $r = 1$). Our primary contribution shows that a three-layer neural network trained via layerwise gradient descent suffices for - complete recovery of the space spanned by the nonlinear features - efficient learning of the target function $f^{\star}=g^{\star}\circ \mathbf{p}$ or transfer learning of $f=g\circ \mathbf{p}$ with a different link function within $\widetilde{\mathcal{O}}(d^4)$ samples and polynomial time.For such hierarchical targets, our result substantially improves the sample complexity ${\Theta}(d^{2p})$ of the kernel methods, demonstrating the power of efficient feature learning. It is important to highlight that our results leverage novel techniques and thus manage to go beyond all prior settings such as single-index and multi-index models as well as models depending just on one nonlinear feature, contributing to a more comprehensive understanding of feature learning in deep learning.

3647. Is Your Model Really A Good Math Reasoner? Evaluating Mathematical Reasoning with Checklist

链接：https://iclr.cc/virtual/2025/poster/28416 abstract： Exceptional mathematical reasoning ability is one of the key features that demonstrate the power of large language models (LLMs). How to comprehensively define and evaluate the mathematical abilities of LLMs, and even reflect the user experience in real-world scenarios, has emerged as a critical issue. Current benchmarks predominantly concentrate on problem-solving capabilities, presenting a substantial risk of model overfitting and fails to accurately measure the genuine mathematical reasoning abilities. In this paper, we argue that if a model really understands a problem, it should be robustly and readily applied across a diverse array of tasks. To this end, we introduce MathCheck, a well-designed checklist for testing task generalization and reasoning robustness, as well as an automatic tool to generate checklists efficiently. MathCheck includes multiple mathematical reasoning tasks and robustness tests to facilitate a comprehensive evaluation of both mathematical reasoning ability and behavior testing. Utilizing MathCheck, we develop MathCheck-GSM and MathCheck-GEO to assess mathematical textual reasoning and multi-modal reasoning capabilities, respectively, serving as upgraded versions of benchmarks including GSM8k, GeoQA, UniGeo, and Geometry3K. We adopt MathCheck-GSM and MathCheck-GEO to evaluate over 26 LLMs and 17 multi-modal LLMs, assessing their comprehensive mathematical reasoning abilities. Our results demonstrate that while frontier LLMs like GPT-4o continue to excel in various abilities on the checklist, many other model families exhibit a significant decline. Further experiments indicate that, compared to traditional math benchmarks, MathCheck better reflects true mathematical abilities and represents mathematical intelligence more linearly, thereby supporting our design. Using MathCheck, we can also efficiently conduct informative behavior analysis to deeply investigate models. Finally, we show that our proposed checklist paradigm can easily extend to other reasoning tasks for their comprehensive evaluation.

3648. Understanding Matrix Function Normalizations in Covariance Pooling through the Lens of Riemannian Geometry

链接：https://iclr.cc/virtual/2025/poster/28270 abstract： Global Covariance Pooling (GCP) has been demonstrated to improve the performance of Deep Neural Networks (DNNs) by exploiting second-order statistics of high-level representations. GCP typically performs classification of the covariance matrices by applying matrix function normalization, such as matrix logarithm or power, followed by a Euclidean classifier. However, covariance matrices inherently lie in a Riemannian manifold, known as the Symmetric Positive Definite (SPD) manifold. The current literature does not provide a satisfactory explanation of why Euclidean classifiers can be applied directly to Riemannian features after the normalization of the matrix power. To mitigate this gap, this paper provides a comprehensive and unified understanding of the matrix logarithm and power from a Riemannian geometry perspective. The underlying mechanism of matrix functions in GCP is interpreted from two perspectives: one based on tangent classifiers (Euclidean classifiers on the tangent space) and the other based on Riemannian classifiers. Via theoretical analysis and empirical validation through extensive experiments on fine-grained and large-scale visual classification datasets, we conclude that the working mechanism of the matrix functions should be attributed to the Riemannian classifiers they implicitly respect. The code is available at https://github.com/GitZH-Chen/RiemGCP.git.

3649. Gyrogroup Batch Normalization

链接：https://iclr.cc/virtual/2025/poster/29016 abstract： Several Riemannian manifolds in machine learning, such as Symmetric Positive Definite (SPD), Grassmann, spherical, and hyperbolic manifolds, have been proven to admit gyro structures, thus enabling a principled and effective extension of Euclidean Deep Neural Networks (DNNs) to manifolds. Inspired by this, this

study introduces a general Riemannian Batch Normalization (RBN) framework on gyrogroups, termed GyroBN. We identify the least requirements to guarantee GyroBN with theoretical control over sample statistics, referred to as \textit{pseudo-reduction} and \textit{gyroisometric gyrations}, which are satisfied by all the existing gyrogroups in machine learning. Besides, our GyroBN incorporates several existing normalization methods, including the one on general Lie groups and different types of RBN on the non-group SPD geometry. Lastly, we instantiate our GyroBN on the Grassmannian and hyperbolic spaces. Experiments on the Grassmannian and hyperbolic networks demonstrate the effectiveness of our GyroBN. The code is available at https://github.com/GitZH-Chen/GyroBN.git.

3650. Convergence of Distributed Adaptive Optimization with Local Updates

链接：https://iclr.cc/virtual/2025/poster/29416 abstract： We study distributed adaptive algorithms with local updates (intermittent communication). Despite the great empirical success of adaptive methods in distributed training of modern machine learning models, the theoretical benefits of local updates within adaptive methods, particularly in terms of reducing communication complexity, have not been fully understood yet. In this paper, for the first time, we prove that \em Local SGD \em with momentum (\em Local \em SGDM) and \em Local \em Adam can outperform their minibatch counterparts in convex and weakly convex settings in certain regimes, respectively. Our analysis relies on a novel technique to prove contraction during local iterations, which is a crucial yet challenging step to show the advantages of local updates, under generalized smoothness assumption and gradient clipping strategy.

3651. Do Egocentric Video-Language Models Truly Understand Hand-Object Interactions?

链接：https://iclr.cc/virtual/2025/poster/29951 abstract： Egocentric video-language pretraining is a crucial step in advancing the understanding of hand-object interactions in first-person scenarios. Despite successes on existing testbeds, we find that current EgoVLMs can be easily misled by simple modifications, such as changing the verbs or nouns in interaction descriptions, with models struggling to distinguish between these changes. This raises the question: "Do EgoVLMs truly understand hand-object interactions?" To address this question, we introduce a benchmark called $\textbf{EgoHOIBench}$, revealing the performance limitation of current egocentric models when confronted with such challenges. We attribute this performance gap to insufficient fine-grained supervision and the greater difficulty EgoVLMs experience in recognizing verbs compared to nouns. To tackle these issues, we propose a novel asymmetric contrastive objective named $\textbf{EgoNCE++}$. For the video-to-text objective, we enhance text supervision by generating negative captions using large language models or leveraging pretrained vocabulary for HOI-related word substitutions. For the text-to-video objective, we focus on preserving an object-centric feature space that clusters video representations based on shared nouns. Extensive experiments demonstrate that EgoNCE++ significantly enhances EgoHOI understanding, leading to improved performance across various EgoVLMs in tasks such as multi-instance retrieval, action recognition, and temporal understanding. Our code is available at https://github.com/xuboshen/EgoNCEpp.

3652. ZeroDiff: Solidified Visual-semantic Correlation in Zero-Shot Learning

链接：https://iclr.cc/virtual/2025/poster/27809 abstract： Zero-shot Learning (ZSL) aims to enable classifiers to identify unseen classes. This is typically achieved by generating visual features for unseen classes based on learned visual-semantic correlations from seen classes. However, most current generative approaches heavily rely on having a sufficient number of samples from seen classes. Our study reveals that a scarcity of seen class samples results in a marked decrease in performance across many generative ZSL techniques. We argue, quantify, and empirically demonstrate that this decline is largely attributable to spurious visual-semantic correlations. To address this issue, we introduce ZeroDiff, an innovative generative framework for ZSL that incorporates diffusion mechanisms and contrastive representations to enhance visual-semantic correlations. ZeroDiff comprises three key components: (1) Diffusion augmentation, which naturally transforms limited data into an expanded set of noised data to mitigate generative model overfitting; (2) Supervised-contrastive (SC)-based representations that dynamically characterize each limited sample to support visual feature generation; and (3) Multiple feature discriminators employing a Wasserstein-distance-based mutual learning approach, evaluating generated features from various perspectives, including pre-defined semantics, SC-based representations, and the diffusion process. Extensive experiments on three popular ZSL benchmarks demonstrate that ZeroDiff not only achieves significant improvements over existing ZSL methods but also maintains robust performance even with scarce training data. Our codes are available at https://github.com/FouriYe/ZeroDiff_ICLR25.

3653. Rodimus*: Breaking the Accuracy-Efficiency Trade-Off with Efficient Attentions

链接：https://iclr.cc/virtual/2025/poster/30177 abstract： Recent advancements in Transformer-based large language models (LLMs) have set new standards in natural language processing. However, the classical softmax attention incurs significant computational costs, leading to a $O(T)$ complexity for per-token generation, where T represents the context length. This work explores reducing LLMs' complexity while maintaining performance by introducing Rodimus and its enhanced version, Rodimus$+$. Rodimus employs an innovative data-dependent tempered selection (DDTS) mechanism within a linear attention-based, purely recurrent framework, achieving significant accuracy while drastically reducing the memory usage typically associated with recurrent models. This method exemplifies semantic compression by maintaining essential input information

with fixed-size hidden states. Building on this, Rodimus$+$ combines Rodimus with the innovative Sliding Window Shared-Key Attention (SW-SKA) in a hybrid approach, effectively leveraging the complementary semantic, token, and head compression techniques. Our experiments demonstrate that Rodimus$+$-1.6B, trained on 1 trillion tokens, achieves superior downstream performance against models trained on more tokens, including Qwen2-1.5B and RWKV6-1.6B, underscoring its potential to redefine the accuracy-efficiency balance in LLMs. Model code and pre-trained checkpoints are open-sourced at https://github.com/codefuse-ai/rodimus.

3654. GrabS: Generative Embodied Agent for 3D Object Segmentation without Scene Supervision

链接：https://iclr.cc/virtual/2025/poster/27837 abstract： We study the hard problem of 3D object segmentation in complex point cloudswithout requiring human labels of 3D scenes for supervision. By relying on thesimilarity of pretrained 2D features or external signals such as motion to group 3Dpoints as objects, existing unsupervised methods are usually limited to identifyingsimple objects like cars or their segmented objects are often inferior due to thelack of objectness in pretrained features. In this paper, we propose a new two-stage pipeline called GrabS. The core concept of our method is to learn generativeand discriminative object-centric priors as a foundation from object datasets in thefirst stage, and then design an embodied agent to learn to discover multiple ob-jects by querying against the pretrained generative priors in the second stage. Weextensively evaluate our method on two real-world datasets and a newly createdsynthetic dataset, demonstrating remarkable segmentation performance, clearlysurpassing all existing unsupervised methods.

3655. Hymba: A Hybrid-head Architecture for Small Language Models

链接：https://iclr.cc/virtual/2025/poster/30655 abstract： We propose Hymba, a family of small language models featuring a hybrid-head parallel architecture that integrates attention mechanisms and state space models (SSMs) within the same layer, offering parallel and complementary processing of the same inputs. In this hybrid-head module, attention heads provide high-resolution recall, while SSM heads facilitate efficient context summarization. Additionally, we introduce learnable meta tokens, which are prepended to prompts to store critical meta information, guiding subsequent tokens and alleviating the “forced-to-attend” burden associated with attention mechanisms. Thanks to the global context summarized by SSMs, the attention heads in our model can be further optimized through cross-layer key-value (KV) sharing and a mix of global and local attention, resulting in a compact cache size without compromising accuracy. Notably, Hymba achieves state-of-the-art performance among small LMs: Our Hymba-1.5B-Base model surpasses all sub-2B public models and even outperforms Llama-3.2-3B, achieving 1.32\% higher average accuracy, an 11.67$\times$ reduction in cache size, and 3.49$\times$ higher throughput.

3656. OBI-Bench: Can LMMs Aid in Study of Ancient Script on Oracle Bones?

链接：https://iclr.cc/virtual/2025/poster/28769 abstract： We introduce OBI-Bench, a holistic benchmark crafted to systematically evaluate large multi-modal models (LMMs) on whole-process oracle bone inscriptions (OBI) processing tasks demanding expert-level domain knowledge and deliberate cognition. OBI-Bench includes 5,523 meticulously collected diverse-sourced images, covering five key domain problems: recognition, rejoining, classification, retrieval, and deciphering. These images span centuries of archaeological findings and years of research by front-line scholars, comprising multi-stage font appearances from excavation to synthesis, such as original oracle bone, inked rubbings, oracle bone fragments, cropped single characters, and handprinted characters. Unlike existing benchmarks, OBI-Bench focuses on advanced visual perception and reasoning with OBI-specific knowledge, challenging LMMs to perform tasks akin to those faced by experts. The evaluation of 6 proprietary LMMs as well as 17 open-source LMMs highlights the substantial challenges and demands posed by OBI-Bench. Even the latest versions of GPT-4o, Gemini 1.5 Pro, and Qwen-VL-Max are still far from public-level humans in some fine-grained perception tasks. However, they perform at a level comparable to untrained humans in deciphering tasks, indicating remarkable capabilities in offering new interpretative perspectives and generating creative guesses. We hope OBI-Bench can facilitate the community to develop domain-specific multi-modal foundation models towards ancient language research and delve deeper to discover and enhance these untapped potentials of LMMs.

3657. QP-SNN: Quantized and Pruned Spiking Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29921 abstract： Brain-inspired Spiking Neural Networks (SNNs) leverage sparse spikes to encode information and operate in an asynchronous event-driven manner, offering a highly energy-efficient paradigm for machine intelligence. However, the current SNN community focuses primarily on performance improvement by developing large-scale models, which limits the applicability of SNNs in resource-limited edge devices. In this paper, we propose a hardware-friendly and lightweight SNN, aimed at effectively deploying high-performance SNN in resource-limited scenarios. Specifically, we first develop a baseline model that integrates uniform quantization and structured pruning, called QP-SNN baseline. While this baseline significantly reduces storage demands and computational costs, it suffers from performance decline. To address this, we conduct an in-depth analysis of the challenges in quantization and pruning that lead to performance degradation and propose solutions to enhance the baseline's performance. For weight quantization, we propose a weight rescaling strategy that utilizes bit width more effectively to enhance the model's representation capability. For structured pruning, we propose a novel pruning criterion using the singular value of spatiotemporal spike activities to enable more accurate removal of redundant kernels. Extensive experiments demonstrate that integrating two proposed methods into the baseline allows QP-

SNN to achieve state-of-the-art performance and efficiency, underscoring its potential for enhancing SNN deployment in edge intelligence computing.

3658. LaMPlace: Learning to Optimize Cross-Stage Metrics in Macro Placement

链接：https://iclr.cc/virtual/2025/poster/29266 abstract： Machine learning techniques have shown great potential in enhancing macro placement, a critical stage in modern chip design.However, existing methods primarily focus on online optimization of intermediate surrogate metrics that are available at the current placement stage, rather than directly targeting the cross-stage metrics---such as the timing performance---that measure the final chip quality.This is mainly because of the high computational costs associated with performing post-placement stages for evaluating such metrics, making the online optimization impractical.Consequently, these optimizations struggle to align with actual performance improvements and can even lead to severe manufacturing issues.To bridge this gap, we propose LaMPlace, which Learns a Mask for optimizing cross-stage metrics in macro placement.Specifically, LaMPlace trains a predictor on offline data to estimate these cross-stage metrics and then leverages the predictor to quickly generate a mask, i.e., a pixel-level feature map that quantifies the impact of placing a macro in each chip grid location on the design metrics.This mask essentially acts as a fast evaluator, enabling placement decisions based on cross-stage metrics rather than intermediate surrogate metrics.Experiments on commonly used benchmarks demonstrate that LaMPlace significantly improves the chip quality across several key design metrics, achieving an average improvement of 9.6\%, notably 43.0\% and 30.4\% in terms of WNS and TNS, respectively, which are two crucial cross-stage metrics that reflect the final chip quality in terms of the timing performance.

3659. Differentiable Integer Linear Programming

链接：https://iclr.cc/virtual/2025/poster/30343 abstract： Machine learning (ML) techniques have shown great potential in generating high-quality solutions for integer linear programs (ILPs).However, existing methods typically rely on a *supervised learning* paradigm, leading to (1) *expensive training cost* due to repeated invocations of traditional solvers to generate training labels, and (2) *plausible yet infeasible solutions* due to the misalignment between the training objective (minimizing prediction loss) and the inference objective (generating high-quality solutions).To tackle this challenge, we propose **DiffILO** (**Diff**erentiable **I**nteger **L**inear Programming **O**ptimization), an *unsupervised learning paradigm for learning to solve ILPs*.Specifically, through a novel probabilistic modeling, DiffILO reformulates ILPs---discrete and constrained optimization problems---into continuous, differentiable (almost everywhere), and unconstrained optimization problems.This reformulation enables DiffILO to simultaneously solve ILPs and train the model via straightforward gradient descent, providing two major advantages.First, it significantly reduces the training cost, as the training process does not need the aid of traditional solvers at all.Second, it facilitates the generation of feasible and high-quality solutions, as the model *learns to solve ILPs* in an end-to-end manner, thus aligning the training and inference objectives.Experiments on commonly used ILP datasets demonstrate that DiffILO not only achieves an average training speedup of 13.2 times compared to supervised methods, but also outperforms them by generating heuristic solutions with significantly higher feasibility ratios and much better solution qualities.

3660. Accelerating Neural ODEs: A Variational Formulation-based Approach

链接：https://iclr.cc/virtual/2025/poster/28017 abstract： Neural Ordinary Differential Equations (Neural ODEs or NODEs) excel at modeling continuous dynamical systems from observational data, especially when the data is irregularly sampled. However, existing training methods predominantly rely on numerical ODE solvers, which are time-consuming and prone to accumulating numerical errors over time due to autoregression. In this work, we propose VF-NODE, a novel approach based on the variational formulation (VF) to accelerate the training of NODEs. Unlike existing training methods, the proposed VF-NODEs implement a series of global integrals, thus evaluating Deep Neural Network (DNN)--based vector fields only at specific observed data points. This strategy drastically reduces the number of function evaluations (NFEs). Moreover, our method eliminates the use of autoregression, thereby reducing error accumulations for modeling dynamical systems. Nevertheless, the VF loss introduces oscillatory terms into the integrals when using the Fourier basis. We incorporate Filon's method to address this issue. To further enhance the performance for noisy and incomplete data, we employ the natural cubic spline regression to estimate a closed-form approximation. We provide a fundamental analysis of how our approach minimizes computational costs. Extensive experiments demonstrate that our approach accelerates NODE training by 10 to 1000 times compared to existing NODE-based methods, while achieving higher or comparable accuracy in dynamical systems. The code is available at https://github.com/ZhaoHongjue/VF-NODE-ICLR2025.

3661. Text2PDE: Latent Diffusion Models for Accessible Physics Simulation

链接：https://iclr.cc/virtual/2025/poster/29869 abstract： Recent advances in deep learning have inspired numerous works on data-driven solutions to partial differential equation (PDE) problems. These neural PDE solvers can often be much faster than their numerical counterparts; however, each presents its unique limitations and generally balances training cost, numerical accuracy, and ease of applicability to different problem setups. To address these limitations, we introduce several methods to apply latent diffusion models to physics simulation. Firstly, we introduce a mesh autoencoder to compress arbitrarily discretized PDE data, allowing for efficient diffusion training across various physics. Furthermore, we investigate full spatiotemporal solution generation to mitigate autoregressive error accumulation. Lastly, we investigate conditioning on initial physical quantities, as well as conditioning solely on a text prompt to introduce text2PDE generation. We show that language can be a compact,

interpretable, and accurate modality for generating physics simulations, paving the way for more usable and accessible PDE solvers. Through experiments on both uniform and structured grids, we show that the proposed approach is competitive with current neural PDE solvers in both accuracy and efficiency, with promising scaling behavior up to $\sim$3 billion parameters. By introducing a scalable, accurate, and usable physics simulator, we hope to bring neural PDE solvers closer to practical use.

3662. Self-Evolving Multi-Agent Collaboration Networks for Software Development

链接：https://iclr.cc/virtual/2025/poster/31011 abstract： LLM-driven multi-agent collaboration (MAC) systems have demonstrated impressive capabilities in automatic software development at the function level. However, their heavy reliance on human design limits their adaptability to the diverse demands of real-world software development.To address this limitation, we introduce EvoMAC, a novel self-evolving paradigm for MAC networks. Inspired by traditional neural network training, EvoMAC obtains text-based environmental feedback by verifying the MAC network's output against a target proxy and leverages a novel textual backpropagation to update the network.To extend coding capabilities beyond function-level tasks to more challenging software-level development, we further propose RSD-Bench, a requirement-oriented software development benchmark, which features complex and diverse software requirements along with automatic evaluation of requirement correctness.Our experiments show that:i) The automatic requirement-aware evaluation in RSD-Bench closely aligns with human evaluations, validating its reliability as a software-level coding benchmark.ii) EvoMAC outperforms previous SOTA methods on both the software-level RSD-Bench and the function-level HumanEval benchmarks, reflecting its superior coding capabilities.

3663. Monte Carlo Planning with Large Language Model for Text-Based Game Agents

链接：https://iclr.cc/virtual/2025/poster/28224 abstract： Text-based games provide valuable environments for language-based autonomous agents. However, planning-then-learning paradigms, such as those combining Monte Carlo Tree Search (MCTS) and reinforcement learning (RL), are notably time-consuming due to extensive iterations. Additionally, these algorithms perform uncertainty-driven exploration but lack language understanding and reasoning abilities.In this paper, we introduce the Monte Carlo planning with Dynamic Memory-guided Large language model (MC-DML) algorithm. MC-DML leverages the language understanding and reasoning capabilities of Large Language Models (LLMs) alongside the exploratory advantages of tree search algorithms. Specifically, we enhance LLMs with in-trial and cross-trial memory mechanisms, enabling them to learn from past experiences and dynamically adjust action evaluations during planning. We conduct experiments on a series of text-based games from the Jericho benchmark. Our results demonstrate that the MC-DML algorithm significantly enhances performance across various games at the initial planning phase, outperforming strong contemporary methods that require multiple iterations. This demonstrates the effectiveness of our algorithm, paving the way for more efficient language-grounded planning in complex environments.

3664. Why RoPE Struggles to Maintain Long-Term Decay in Long Sequences?

链接：https://iclr.cc/virtual/2025/poster/31353 abstract： Rotary Position Embedding (RoPE) improves upon traditional positional encodings but struggles with long-term decay in contexts exceeding its training length, limiting the model's generalization to longer sequences. Our experiments suggest that this issue may stem from a high proportion of obtuse angles on the complex plane between the linear transformations of query and key embeddings.

3665. TidalDecode: Fast and Accurate LLM Decoding with Position Persistent Sparse Attention

链接：https://iclr.cc/virtual/2025/poster/30389 abstract： Large language models (LLMs) have driven significant advancements across diverse NLP tasks, with long-context models gaining prominence for handling extended inputs. However, the expanding key-value (KV) cache size required by Transformer architectures intensifies the memory constraints, particularly during the decoding phase, creating a significant bottleneck. Existing sparse attention mechanisms designed to address this bottleneck have two limitations: (1) they often fail to reliably identify the most relevant tokens for attention, and (2) they overlook the spatial coherence of token selection across consecutive Transformer layers, which can lead to performance degradation and substantial overhead in token selection. This paper introduces TidalDecode, a simple yet effective algorithm and system for fast and accurate LLM decoding through position persistent sparse attention. TidalDecode leverages the spatial coherence of tokens selected by existing sparse attention methods and introduces a few token selection layers that perform full attention to identify the tokens with the highest attention scores, while all other layers perform sparse attention with the pre-selected tokens. This design enables TidalDecode to substantially reduce the overhead of token selection for sparse attention without sacrificing the quality of the generated results. Evaluation on a diverse set of LLMs and tasks shows that TidalDecode closely matches the generative performance of full attention methods while reducing the LLM decoding latency by up to $2.1\times$.

3666. Solving New Tasks by Adapting Internet Video Knowledge

链接：https://iclr.cc/virtual/2025/poster/28326 abstract： Video generative models demonstrate great promise in robotics by serving as visual planners or as policy supervisors. When pretrained on internet-scale data, such video models intimately understand alignment with natural language, and can thus facilitate generalization to novel downstream behavior through text-conditioning. However, they may not be sensitive to the specificities of the particular environment the agent inhabits. On the other hand, training video models on in-domain examples of robotic behavior naturally encodes environment-specific intricacies, but the scale of available demonstrations may not be sufficient to support generalization to unseen tasks via natural language specification. In this work, we investigate different adaptation techniques that integrate in-domain information with large-scale pretrained video models, and explore the extent to which they enable novel text-conditioned generalization for robotic tasks, while also considering their independent data and resource considerations. We successfully demonstrate across robotic environments that adapting powerful video models with small scales of example data can successfully facilitate generalization to novel behaviors. In particular, we present a novel adaptation strategy, termed Inverse Probabilistic Adaptation, that not only consistently achieves strong generalization performance across robotic tasks and settings, but also exhibits robustness to the quality of adaptation data, successfully solving novel tasks even when only suboptimal in-domain demonstrations are available.

3667. Have the VLMs Lost Confidence? A Study of Sycophancy in VLMs

链接：https://iclr.cc/virtual/2025/poster/30427 abstract： In the study of LLMs, sycophancy represents a prevalent hallucination that poses significant challenges to these models. Specifically, LLMs often fail to adhere to original correct responses, instead blindly agreeing with users' opinions, even when those opinions are incorrect or malicious. However, research on sycophancy in visual language models (VLMs) has been scarce. In this work, we extend the exploration of sycophancy from LLMs to VLMs, introducing the MM-SY benchmark to evaluate this phenomenon. We present evaluation results from multiple representative models, addressing the gap in sycophancy research for VLMs. To mitigate sycophancy, we propose a synthetic dataset for training and employ methods based on prompts, supervised fine-tuning, and DPO. Our experiments demonstrate that these methods effectively alleviate sycophancy in VLMs. Additionally, we probe VLMs to assess the semantic impact of sycophancy and analyze the attention distribution of visual tokens. Our findings indicate that the ability to prevent sycophancy is predominantly observed in higher layers of the model. The lack of attention to image knowledge in these higher layers may contribute to sycophancy, and enhancing image attention at high layers proves beneficial in mitigating this issue.

3668. Intervening Anchor Token: Decoding Strategy in Alleviating Hallucinations for MLLMs

链接：https://iclr.cc/virtual/2025/poster/27678 abstract： Multimodal large language models (MLLMs) offer a powerful mechanism for interpreting visual information. However, they often suffer from hallucinations, which impede the real-world usage of these models. Existing methods attempt to alleviate this issue by designing special decoding strategies that penalize the summary tokens. However, these methods lack analysis of the relationship between hallucination and summarization mechanism of LLMs. Interestingly, we find that penalizing summary tokens is not necessary: merely intervening the query-key parameters variance, without costing extra inference time, still alleviates hallucinations. Specifically, we explore the causes of hallucinations by analyzing localized self-attention patterns called ``anchor'' tokens and define the attention localization degree of the model as token propagation probabilities. Our analysis reveals that over-propagation of anchor tokens occurs when the distribution of eigenvalues of the query and key matrices has a non-zero mean and a polarized variance, leading to excessive dependence on anchor tokens while neglecting vision information and describes the image content with hallucination. Based on the observation, we propose a versatile plug-and-play decoding strategy, Dynamic Token Propagation Mechanism (TAME), to alleviate excessive propagation by dynamically intervening the eigenspectrum variance of the attention weight, thereby alleviating hallucinations without relying on complex decoding strategies. Extensive experiments reveal a correlation between the eigenspectrum and hallucinations across various MLLMs, and show that TAME reduces the percentage of hallucinated objects.

3669. Neuralized Markov Random Field for Interaction-Aware Stochastic Human Trajectory Prediction

链接：https://iclr.cc/virtual/2025/poster/28222 abstract： Interactive human motions and the continuously changing nature of intentions pose significant challenges for human trajectory prediction. In this paper, we present a neuralized Markov random field (MRF)-based motion evolution method for probabilistic interaction-aware human trajectory prediction. We use MRF to model each agent's motion and the resulting crowd interactions over time, hence is robust against noisy observations and enables group reasoning. We approximate the modeled distribution using two conditional variational autoencoders (CVAEs) for efficient learning and inference. Our proposed method achieves state-of-the-art performance on ADE/FDE metrics across two dataset categories: overhead datasets ETH/UCY, SDD, and NBA, and ego-centric JRDB. Furthermore, our approach allows for real-time stochastic inference in bustling environments, making it well-suited for a 30FPS video setting. We open-source our codes at: https://github.com/AdaCompNUS/NMRF_TrajectoryPrediction.git

3670. Post-hoc Reward Calibration: A Case Study on Length Bias

链接：https://iclr.cc/virtual/2025/poster/30144 abstract： Reinforcement Learning from Human Feedback aligns the outputs of Large Language Models with human values and preferences. Central to this process is the reward model (RM), which translates human feedback into training signals for optimising LLM behaviour. However, RMs can develop biases by exploiting spurious correlations in their training data, such as favouring outputs based on length orstyle rather than true quality. These biases can

lead to incorrect output rankings, sub-optimal model evaluations, and the amplification of undesirable behaviours in LLMs alignment. This paper addresses the challenge of correcting such biases without additional data and training, introducing the concept of Post-hoc Reward Calibration. We first propose to use local average reward to estimate the bias termand, thus, remove it to approximate the underlying true reward. We then extend the approach to a more general and robust form with the Locally Weighted Regression. Focusing on the prevalent length bias, we validate our proposed approaches across three experimental settings, demonstrating consistent improvements: (1) a 3.11 average performance gain across 33 reward models on the RewardBenchdataset; (2) improved agreement of RM produced rankings with GPT-4 evaluations and human preferences based on the AlpacaEval benchmark; and (3) improved Length-Controlled win rate (Dubois et al., 2024) of the RLHF process in multiple LLM–RM combinations. According to our experiments, our method is computationally efficient and generalisable to other types of bias and RMs, offering a scalable and robust solution for mitigating biases in LLM alignment and evaluation.

3671. Layerwise Recurrent Router for Mixture-of-Experts

链接：https://iclr.cc/virtual/2025/poster/28923 abstract： The scaling of large language models (LLMs) has revolutionized their capabilities in various tasks, yet this growth must be matched with efficient computational strategies. The Mixture-of-Experts (MoE) architecture stands out for its ability to scale model size without significantly increasing training costs. Despite their advantages, current MoE models often display parameter inefficiency. For instance, a pre-trained MoE-based LLM with 52 billion parameters might perform comparably to a standard model with 6.7 billion. Being a crucial part of MoE, current routers in different layers independently assign tokens without leveraging historical routing information, potentially leading to suboptimal token-expert combinations and the parameter inefficiency problem.To alleviate this issue, we introduce the Layerwise Recurrent Router for Mixture-of-Experts (RMoE). RMoE leverages a Gated Recurrent Unit (GRU) to establish dependencies between routing decisions across consecutive layers.Such layerwise recurrence can be efficiently parallelly computed for input tokens and introduces negotiable costs.Our extensive empirical evaluations demonstrate that RMoE-based language models consistently outperform a spectrum of baseline models. Furthermore, RMoE integrates a novel computation stage orthogonal to existing methods, allowing seamless compatibility with other MoE architectures. Our analyses attribute RMoE's gains to its effective cross-layer information sharing, which also improves expert selection and diversity.

3672. Explaining Modern Gated-Linear RNNs via a Unified Implicit Attention Formulation

链接：https://iclr.cc/virtual/2025/poster/27821 abstract： Recent advances in efficient sequence modeling have led to attention-free layers, such as Mamba, RWKV, and various gated RNNs, all featuring sub-quadratic complexity in sequence length and excellent scaling properties, enabling the construction of a new type of foundation models. In this paper, we present a unified view of these models, formulating such layers as implicit causal self-attention layers. The formulation includes most of their sub-components and is not limited to a specific part of the architecture. The framework compares the underlying mechanisms on similar grounds for different layers and provides a direct means for applying explainability methods. Our experiments show that our attention matrices and attribution method outperform an alternative and a more limited formulation that was recently proposed for Mamba. For the other architectures for which our method is the first to provide such a view, our method is effective and competitive in the relevant metrics compared to the results obtained by state-of-the-art Transformer explainability methods. Our code is publicly available.

3673. DeciMamba: Exploring the Length Extrapolation Potential of Mamba

链接：https://iclr.cc/virtual/2025/poster/28694 abstract： Long-range sequence processing poses a significant challenge for Transformers due to their quadratic complexity in input length. A promising alternative is Mamba, which demonstrates high performance and achieves Transformer-level capabilities while requiring substantially fewer computational resources. In this paper we explore the length-generalization capabilities of Mamba, which we find to be relatively limited. Through a series of visualizations and analyses we identify that the limitations arise from a restricted effective receptive field, dictated by the sequence length used during training. To address this constraint, we introduce DeciMamba, a context-extension method specifically designed for Mamba. This mechanism, built on top of a hidden filtering mechanism embedded within the S6 layer, enables the trained model to extrapolate well even without additional training. Empirical experiments over real-world long-range NLP tasks show that DeciMamba can extrapolate to context lengths that are significantly longer than the ones seen during training, while enjoying faster inference. We will release our code and models.

3674. Internet of Agents: Weaving a Web of Heterogeneous Agents for Collaborative Intelligence

链接：https://iclr.cc/virtual/2025/poster/28382 abstract： The rapid advancement of large language models (LLMs) has paved the way for the development of highly capable autonomous agents. However, existing multi-agent frameworks often struggle with integrating diverse capable third-party agents due to reliance on agents defined within their own ecosystems. They also face challenges in simulating distributed environments, as most frameworks are limited to single-device setups. Furthermore, these frameworks often rely on hard-coded communication pipelines, limiting their adaptability to dynamic task requirements. Inspired by the concept of the Internet, we propose the Internet of Agents (IoA), a novel framework that addresses these limitations by providing a flexible and scalable platform for LLM-based multi-agent collaboration. IoA introduces an agent integration protocol, an instant-messaging-like architecture design, and dynamic mechanisms for agent teaming and conversation flow control.

Through extensive experiments on general assistant tasks, embodied AI tasks, and retrieval-augmented generation benchmarks, we demonstrate that IoA consistently outperforms state-of-the-art baselines, showcasing its ability to facilitate effective collaboration among heterogeneous agents. IoA represents a step towards linking diverse agents in an Internet-like environment, where agents can seamlessly collaborate to achieve greater intelligence and capabilities. We will release our code to facilitate further research.

3675. On the Byzantine-Resilience of Distillation-Based Federated Learning

链接：https://iclr.cc/virtual/2025/poster/28341 abstract： Federated Learning (FL) algorithms using Knowledge Distillation (KD) have received increasing attention due to their favorable properties with respect to privacy, non-i.i.d. data and communication cost. These methods depart from transmitting model parameters and instead communicate information about a learning task by sharing predictions on a public dataset. In this work, we study the performance of such approaches in the byzantine setting, where a subset of the clients act in an adversarial manner aiming to disrupt the learning process. We show that KD-based FL algorithms are remarkably resilient and analyze how byzantine clients can influence the learning process. Based on these insights, we introduce two new byzantine attacks and demonstrate their ability to break existing byzantine-resilient methods. Additionally, we propose a novel defence method which enhances the byzantine resilience of KD-based FL algorithms. Finally, we provide a general framework to obfuscate attacks, making them significantly harder to detect, thereby improving their effectiveness.

3676. Preserving Diversity in Supervised Fine-Tuning of Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29878 abstract： Large Language Models (LLMs) typically rely on Supervised Fine-Tuning (SFT) to specialize in downstream tasks, with the Cross Entropy (CE) loss being the de facto choice. However, CE maximizes the likelihood of observed data without accounting for alternative possibilities. As such, CE usually leads to reduced diversity in the model's outputs, which hinders further development that requires sampling to explore better responses. To address this limitation, this paper introduces a new game-theoretic formulation for SFT. In this framework, an auxiliary variable is introduced to regulate the learning process. We prove that the proposed game-theoretic approach connects to the problem of reverse KL minimization with entropy regularization. This regularization prevents over-memorization of training data and promotes output diversity. To implement this framework, we develop GEM, a new training algorithm that is computationally efficient as CE by leveraging some unique properties of LLMs. Empirical studies of pre-trained models from 3B to 70B parameters show that GEM achieves comparable downstream performance to CE while significantly enhancing output diversity. This increased diversity translates to performance gains in test-time compute scaling for chat and code generation tasks. Moreover, we observe that preserving output diversity has the added benefit of mitigating forgetting, as maintaining diverse outputs encourages models to retain pre-trained knowledge throughout the training process.

3677. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention

链接：https://iclr.cc/virtual/2025/poster/30556 abstract： Large Vision-Language Models (LVLMs) exhibit impressive capabilities in complex visual tasks but are prone to hallucination, especially in open-ended generation tasks. This paper explores why LVLMs tend to hallucinate and how to mitigate it. First, we conduct causal mediation analysis through counterfactual edits on specific modules in LVLMs. Our results disclose that Multi-Head Attention (MHA) modules contribute more to the probability of generating hallucination words than multi-layer perceptron modules. We then identify specific heads that are responsible for hallucination, referred to as hallucination heads. Second, we examine the behavior of hallucination heads. We find that they are concentrated in the middle and deeper layers, displaying a strong attention bias toward text tokens. Further, we show that the attention patterns of certain hallucination heads exhibit greater similarity to the base language model and change slowly during the instruction tuning process. Finally, we propose two simple yet effective methods to mitigate hallucination: one is training-free and can be applied directly during decoding, while the other involves fine-tuning. Both methods are targeted for hallucination heads to reduce their reliance on text tokens. Notably, our methods achieve up to 1.7x reduction in hallucination rate for the LLaVA-v1.5-7B model in COCO captioning task, outperforming existing baselines. Overall, our findings suggest that hallucinations in LVLMs are likely to stem from certain modules, and targeted interventions can effectively mitigate these issues.

3678. Adam-mini: Use Fewer Learning Rates To Gain More

链接：https://iclr.cc/virtual/2025/poster/28710 abstract： We propose Adam-mini, an optimizer that achieves on-par or better performance than AdamW with 50% less memory footprint. Adam-mini reduces memory by cutting down the learning rate resources in Adam (i.e., $1/\sqrt{v}$). By delving into the Hessian structure of neural nets, we find Adam’s v might not function at its full potential as effectively as we expected. We find that ≥ 99.9% of these learning rates in v could be harmlessly removed if we (1) carefully partition the parameters into blocks following our proposed principle on Hessian structure; (2) assign a single but good learning rate to each parameter block. We then provide one simple way to find good learning rates and propose Adam-mini. Empirically, we verify that Adam-mini performs on par or better than AdamW on various language models sized from 39M to 13B for pre-training, supervised fine-tuning, and RLHF. The reduced memory footprint of Adam-mini also alleviates communication overheads among GPUs, thereby increasing throughput. For instance, Adam-mini achieves 49.6%

higher throughput than AdamW when pre-training Llama 2-7B on $2\times$ A800-80GB GPUs, which saves 33% wall-clock time for pre-training.

3679. Learning to Adapt Frozen CLIP for Few-Shot Test-Time Domain Adaptation

链接：https://iclr.cc/virtual/2025/poster/29558 abstract： Few-shot Test-Time Domain Adaptation focuses on adapting a model at test time to a specific domain using only a few unlabeled examples, addressing domain shift. Prior methods leverage CLIP's strong out-of-distribution (OOD) abilities by generating domain-specific prompts to guide its generalized, frozen features. However, since downstream datasets are not explicitly seen by CLIP, solely depending on the feature space knowledge is constrained by CLIP's prior knowledge. Notably, when using a less robust backbone like ViT-B/16, performance significantly drops on challenging real-world benchmarks. Departing from the state-of-the-art of inheriting the intrinsic OOD capability of CLIP, this work introduces learning directly on the input space to complement the dataset-specific knowledge for frozen CLIP. Specifically, an independent side branch is attached in parallel with CLIP and enforced to learn exclusive knowledge via revert attention. To better capture the dataset-specific label semantics for downstream adaptation, we propose to enhance the inter-dispersion among text features via greedy text ensemble and refinement. The text and visual features are then progressively fused in a domain-aware manner by a generated domain prompt to adapt toward a specific domain. Extensive experiments show our method's superiority on 5 large-scale benchmarks (WILDS and DomainNet), notably improving over smaller networks like ViT-B/16 with gains of \textbf{+5.1} in F1 for iWildCam and \textbf{+3.1\%} in WC Acc for FMoW. \href{https://github.com/chi-chi-zx/L2C}{Our Code: L2C}

3680. As Simple as Fine-tuning: LLM Alignment via Bidirectional Negative Feedback Loss

链接：https://iclr.cc/virtual/2025/poster/28849 abstract： Direct Preference Optimization (DPO) has emerged as a more computationally efficient alternative to Reinforcement Learning from Human Feedback (RLHF) with Proximal Policy Optimization (PPO), eliminating the need for reward models and online sampling. Despite these benefits, DPO and its variants remain sensitive to hyper-parameters and prone to instability, particularly on mathematical datasets. We argue that these issues arise from the unidirectional likelihood-derivative negative feedback inherent in the log-likelihood loss function.To address this, we propose a novel LLM alignment loss that establishes a stable Bidirectional Negative Feedback (BNF) during optimization. Our proposed BNF loss eliminates the need for pairwise contrastive losses and does not require any extra tunable hyper-parameters or pairwise preference data, streamlining the alignment pipeline to be as simple as supervised fine-tuning.We conduct extensive experiments across two challenging QA benchmarks and four reasoning benchmarks. The experimental results show that BNF achieves comparable performance to the best methods on QA benchmarks, while its performance decrease on the four reasoning benchmarks is significantly lower compared to the best methods, thus striking a better balance between value alignment and reasoning ability. In addition, we further validate the performance of BNF on non-pairwise datasets, and conduct in-depth analysis of log-likelihood and logit shifts across different preference optimization methods.We will release all the source code, checkpoints, and datasets on GitHub.

3681. Multiple Heads are Better than One: Mixture of Modality Knowledge Experts for Entity Representation Learning

链接：https://iclr.cc/virtual/2025/poster/27966 abstract： Learning high-quality multi-modal entity representations is an important goal of multi-modal knowledge graph (MMKG) representation learning, which can en- hance reasoning tasks within the MMKGs, such as MMKG completion (MMKGC). The main challenge is to collaboratively model the structural information concealed in massive triples and the multi-modal features of the entities. Existing methods focus on crafting elegant entity-wise multi-modal fusion strategies, yet they over- look the utilization of multi-perspective features concealed within the modalities under diverse relational contexts. To address this issue, we introduce a novel framework with Mixture of Modality Knowledge experts (MOMOK for short) to learn adaptive multi-modal entity representations for better MMKGC. We design relation-guided modality knowledge experts to acquire relation-aware modality embeddings and integrate the predictions from multi-modalities to achieve joint decisions. Additionally, we disentangle the experts by minimizing their mutual information. Experiments on four public MMKG benchmarks demonstrate the outstanding performance of MOMOK under complex scenarios. Our code and data are available at https://github.com/zjukg/MoMoK.

3682. LoRA3D: Low-Rank Self-Calibration of 3D Geometric Foundation models

链接：https://iclr.cc/virtual/2025/poster/29996 abstract： Emerging 3D geometric foundation models, such as DUSt3R, offer a promising approach for in-the-wild 3D vision tasks.However, due to the high-dimensional nature of the problem space and scarcity of high-quality 3D data,these pre-trained models still struggle to generalize to many challenging circumstances,such as limited view overlap or low lighting.To address this, we propose LoRA3D, an efficient self-calibration pipeline to specialize the pre-trained models to target scenes using their own multi-view predictions.Taking sparse RGB images as input, we leverage robust optimization techniques to refine multi-view predictions and align them into a global coordinate frame.In particular, we incorporate prediction confidence into the geometric optimization process, automatically re-weighting the confidence to better

reflect point estimation accuracy. We use the calibrated confidence to generate high-quality pseudo labels for the calibrating views and fine-tune the models using low-rank adaptation (LoRA) on the pseudo-labeled data.Our method does not require any external priors or manual labels. It completes the self-calibration process on a single standard GPU within just 5 minutes.Each low-rank adapter requires only 18MB of storage. We evaluated our method on more than 160 scenes from the Replica, TUM and Waymo Open datasets,achieving up to 88\% performance improvement on 3D reconstruction, multi-view pose estimation and novel-view rendering.For more details, please visit our project page at https://520xyxyzq.github.io/lora3d/.

3683. Combatting Dimensional Collapse in LLM Pre-Training Data via Submodular File Selection

链接：https://iclr.cc/virtual/2025/poster/28897 abstract： Selecting high-quality pre-training data for large language models (LLMs) is crucial for enhancing their overall performance under limited computation budget, improving both training and sample efficiency. Recent advancements in file selection primarily rely on using an existing or trained proxy model to assess the similarity of samples to a target domain, such as high quality sources BookCorpus and Wikipedia. However, upon revisiting these methods, the domain-similarity selection criteria demonstrates a diversity dilemma, i.e. dimensional collapse in the feature space, improving performance on the domain-related tasks but causing severe degradation on generic performance.To prevent collapse and enhance diversity, we propose a DiverSified File selection algorithm (DiSF), which selects the most decorrelated text files in the feature space. We approach this with a classical greedy algorithm to achieve more uniform eigenvalues in the feature covariance matrix of the selected texts, analyzing its approximation to the optimal solution under a formulation of γ-weakly submodular optimization problem. Empirically, we establish a benchmark and conduct extensive experiments on the TinyLlama architecture with models from 120M to 1.1B parameters. Evaluating across nine tasks from the Harness framework, DiSF demonstrates a significant improvement on overall performance. Specifically, DiSF saves 98.5\% of 590M training files in SlimPajama, outperforming the full-data pre-training within a 50B training budget, and achieving about 1.5x training efficiency and 5x data efficiency. Source codeis available at: https://github.com/MediaBrain-SJTU/DiSF.git.

3684. Aligning Generative Denoising with Discriminative Objectives Unleashes Diffusion for Visual Perception

链接：https://iclr.cc/virtual/2025/poster/28201 abstract： With success in image generation, generative diffusion models are increasingly adopted for discriminative scenarios because generating pixels is a unified and natural perception interface. Although directly re-purposing their generative denoising process has established promising progress in specialist (e.g., depth estimation) and generalist models, the inherent gaps between a generative process and discriminative objectives are rarely investigated. For instance, generative models can tolerate deviations at intermediate sampling steps as long as the final distribution is reasonable, while discriminative tasks with rigorous ground truth for evaluation are sensitive to such errors. Without mitigating such gaps, diffusion for perception still struggles on tasks represented by multi-modal understanding (e.g., referring image segmentation). Motivated by these challenges, we analyze and improve the alignment between the generative diffusion process and perception objectives centering around the key observation: how perception quality evolves with the denoising process. (1) Notably, earlier denoising steps contribute more than later steps, necessitating a tailored learning objective for training: loss functions should reflect varied contributions of timesteps for each perception task. (2) Perception quality drops unexpectedly at later denoising steps, revealing the sensitiveness of perception to training-denoising distribution shift. We introduce diffusion-tailored data augmentation to simulate such drift in the training data. (3) We suggest a novel perspective to the long-standing question: why should a generative process be useful for discriminative tasks - interactivity. The denoising process can be leveraged as a controllable user interface adapting to users' correctional prompts and conducting multi-round interaction in an agentic workflow. Collectively, our insights enhance multiple generative diffusion-based perception models without architectural changes: state-of-the-art diffusion-based depth estimator, previously underplayed referring image segmentation models, and perception generalists. Our code is available at https://github.com/ziqipang/ADDP.

3685. Eliminating Position Bias of Language Models: A Mechanistic Approach

链接：https://iclr.cc/virtual/2025/poster/28841 abstract： Position bias has proven to be a prevalent issue of modern language models (LMs), where the models prioritize content based on its position within the given context. This bias often leads to unexpected model failures and hurts performance, robustness, and reliability across various applications. A simple mechanistic analysis attributes the position bias to two components employed in nearly all state-of-the-art LMs: causal attention and position embedding. Based on the analyses, we propose to **eliminate** position bias (e.g., different retrieved documents' orders in QA affect performance) with a **training-free zero-shot** approach. Our method changes the causal attention to bidirectional attention between documents and utilizes model attention values to decide the relative orders of documents instead of using the order provided in input prompts, therefore enabling Position-INvariant inferencE (PINE) at the document level. By eliminating position bias, models achieve better performance and reliability in downstream tasks, including LM-as-a-judge, retrieval-augmented QA, molecule generation, and math reasoning. Notably, PINE is especially useful when adapting LMs for evaluating reasoning pairs: it consistently provides 8 to 10 percentage points performance gains, making Llama-3-70B-Instruct perform even better than GPT-4-0125-preview and GPT-4o-2024-08-06 on the RewardBench reasoning set.

3686. Query-based Knowledge Transfer for Heterogeneous Learning

Environments

链接：https://iclr.cc/virtual/2025/poster/29312 abstract： Decentralized collaborative learning under data heterogeneity and privacy constraints has rapidly advanced. However, existing solutions like federated learning, ensembles, and transfer learning, often fail to adequately serve the unique needs of clients, especially when local data representation is limited. To address this issue, we propose a novel framework called Query-based Knowledge Transfer (QKT) that enables tailored knowledge acquisition to fulfill specific client needs without direct data exchange. It employs a data-free masking strategy to facilitate the communication-efficient query-focused knowledge transformation while refining task-specific parameters to mitigate knowledge interference and forgetting. Our experiments, conducted on both standard and clinical benchmarks, show that QKT significantly outperforms existing collaborative learning methods by an average of 20.91% points in single-class query settings and an average of 14.32% points in multi-class query scenarios.Further analysis and ablation studies reveal that QKT effectively balances the learning of new and existing knowledge, showing strong potential for its application in decentralized learning.

3687. SEBRA : Debiasing through Self-Guided Bias Ranking

链接：https://iclr.cc/virtual/2025/poster/29907 abstract： Ranking samples by fine-grained estimates of spuriosity (the degree to which spurious cues are present) has recently been shown to significantly benefit bias mitigation, over the traditional binary biased-vs-unbiased partitioning of train sets. However, this spuriousity ranking comes with the requirement of human supervision. In this paper, we propose a debiasing framework based on our novel Self-Guided Bias Ranking (Sebra), that mitigates biases via an automatic ranking of data points by spuriosity within their respective classes. Sebra leverages a key local symmetry in Empirical Risk Minimization (ERM) training -- the ease of learning a sample via ERM inversely correlates with its spuriousity; the fewer spurious correlations a sample exhibits, the harder it is to learn, and vice versa. However, globally across iterations, ERM tends to deviate from this symmetry. Sebra dynamically steers ERM to correct this deviation, facilitating the sequential learning of attributes in increasing order of difficulty, ie, decreasing order of spuriosity. As a result, the sequence in which Sebra learns samples naturally provides spuriousity rankings. We use the resulting fine-grained bias characterization in a contrastive learning framework to mitigate biases from multiple sources. Extensive experiments show that Sebra consistently outperforms previous state-of-the-art unsupervised debiasing techniques across multiple standard benchmarks, including UrbanCars, BAR, and CelebA.

3688. STAMP: Scalable Task- And Model-agnostic Collaborative Perception

链接：https://iclr.cc/virtual/2025/poster/30771 abstract： Perception is a crucial component of autonomous driving systems. However, single-agent setups often face limitations due to sensor constraints, especially under challenging conditions like severe occlusion, adverse weather, and long-range object detection. Multi-agent collaborative perception (CP) offers a promising solution that enables communication and information sharing between connected vehicles. Yet, the heterogeneity among agents—in terms of sensors, models, and tasks—significantly hinders effective and efficient cross-agent collaboration. To address these challenges, we propose STAMP, a scalable task- and model-agnostic collaborative perception framework tailored for heterogeneous agents. STAMP utilizes lightweight adapter-reverter pairs to transform Bird's Eye View (BEV) features between agent-specific domains and a shared protocol domain, facilitating efficient feature sharing and fusion while minimizing computational overhead. Moreover, our approach enhances scalability, preserves model security, and accommodates a diverse range of agents. Extensive experiments on both simulated (OPV2V) and real-world (V2V4Real) datasets demonstrate that STAMP achieves comparable or superior accuracy to state-of-the-art models with significantly reduced computational costs. As the first-of-its-kind task- and model-agnostic collaborative perception framework, STAMP aims to advance research in scalable and secure mobility systems, bringing us closer to Level 5 autonomy. Our project page is at https://xiangbogaobarry.github.io/STAMP and the code is available at https://github.com/taco-group/STAMP.

3689. A Statistical Approach for Controlled Training Data Detection

链接：https://iclr.cc/virtual/2025/poster/29318 abstract： Detecting training data for large language models (LLMs) is receiving growing attention, especially in applications requiring high reliability. While numerous efforts have been made to address this issue, they typically focus on accuracy without ensuring controllable results.To fill this gap, we propose Knockoff Inference-based Training data Detector (KTD), a novel method that achieves rigorous false discovery rate (FDR) control in training data detection. Specifically, KTD generates synthetic knockoff samples that seamlessly replace original data points without compromising contextual integrity. A novel knockoff statistic, which incorporates multiple knockoff draws, is then calculated to ensure FDR control while maintaining high power. Our theoretical analysis demonstrates KTD's asymptotic optimality in terms of FDR control and power. Empirical experiments on real-world datasets such as WikiMIA, XSum and Real Time BBC News further validate KTD's superior performance compared to existing methods.

3690. GS-CPR: Efficient Camera Pose Refinement via 3D Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/28467 abstract： We leverage 3D Gaussian Splatting (3DGS) as a scene representation and propose a novel test-time camera pose refinement (CPR) framework, GS-CPR. This framework enhances the localization accuracy of state-of-the-art absolute pose regression and scene coordinate regression methods. The 3DGS model renders high-quality synthetic images and depth maps to facilitate the establishment of 2D-3D correspondences. GS-CPR obviates the need for training feature extractors or descriptors by operating directly on RGB images, utilizing the 3D foundation model, MASt3R, for precise 2D matching. To improve the robustness of our model in challenging outdoor

environments, we incorporate an exposure-adaptive module within the 3DGS framework. Consequently, GS-CPR enables efficient one-shot pose refinement given a single RGB query and a coarse initial pose estimation. Our proposed approach surpasses leading NeRF-based optimization methods in both accuracy and runtime across indoor and outdoor visual localization benchmarks, achieving new state-of-the-art accuracy on two indoor datasets.

3691. Automatic Curriculum Expert Iteration for Reliable LLM Reasoning

链接：https://iclr.cc/virtual/2025/poster/31053 abstract： Hallucinations (i.e., generating plausible but inaccurate content) and laziness (i.e. excessive refusals or defaulting to "I don't know") persist as major challenges in LLM reasoning. Current efforts to reduce hallucinations primarily focus on factual errors in knowledge-grounded tasks, often neglecting hallucinations related to faulty reasoning. Meanwhile, some approaches render LLMs overly conservative, limiting their problem-solving capabilities. To mitigate hallucination and laziness in reasoning tasks, we propose Automatic Curriculum Expert Iteration (Auto-CEI) to enhance LLM reasoning and align responses to the model's capabilities--assertively answering within its limits and declining when tasks exceed them. In our method, Expert Iteration explores the reasoning trajectories near the LLM policy, guiding incorrect paths back on track to reduce compounding errors and improve robustness; it also promotes appropriate "I don't know" responses after sufficient reasoning attempts. The curriculum automatically adjusts rewards, incentivizing extended reasoning before acknowledging incapability, thereby pushing the limits of LLM reasoning and aligning its behaviour with these limits. We compare Auto-CEI with various SOTA baselines across logical reasoning, mathematics, and planning tasks, where Auto-CEI achieves superior alignment by effectively balancing assertiveness and conservativeness.

3692. Advancing Mathematical Reasoning in Language Models: The Impact of Problem-Solving Data, Data Synthesis Methods, and Training Stages

链接：https://iclr.cc/virtual/2025/poster/30252 abstract： Mathematical reasoning remains a challenging area for large language models (LLMs), prompting the development of math-specific LLMs such as LLEMMA, DeepSeekMath, and Qwen2-Math, among others. These models typically follow a two-stage training paradigm: pre-training with math-related corpora and post-training with problem datasets for supervised fine-tuning (SFT). Despite these efforts, the improvements in mathematical reasoning achieved through continued pre-training (CPT) are often less significant compared to those obtained via SFT. This study addresses this discrepancy by exploring alternative strategies during the pre-training phase, focusing on the use of problem-solving data over general mathematical corpora.We investigate three primary research questions: (1) Can problem-solving data enhance the model's mathematical reasoning capabilities more effectively than general mathematical corpora during CPT? (2) Are synthetic data from the same source equally effective, and which synthesis methods are most efficient? (3) How do the capabilities developed from the same problem-solving data differ between the CPT and SFT stages, and what factors contribute to these differences?Our findings indicate that problem-solving data significantly enhances the model's mathematical capabilities compared to general mathematical corpora. We also identify effective data synthesis methods, demonstrating that the tutorship amplification synthesis method achieves the best performance. Furthermore, while SFT facilitates instruction-following abilities, it underperforms compared to CPT with the same data, which can be partially attributed to its poor learning capacity for more challenging problem-solving data. These insights provide valuable guidance for optimizing the mathematical reasoning capabilities of LLMs, culminating in our development of a powerful mathematical base model called MathGPT-8B.

3693. Synthetic continued pretraining

链接：https://iclr.cc/virtual/2025/poster/31270 abstract： Pretraining on large-scale, unstructured internet text enables language models to acquire a significant amount of world knowledge.However, this knowledge acquisition is data-inefficient---to learn a fact, models must be trained on hundreds to thousands of diverse representations of it.This poses a challenge when adapting a pretrained model to a small corpus of domain-specific documents, where each fact may appear rarely or only once.We propose to bridge this gap with synthetic continued pretraining: using the small domain-specific corpus to synthesize a large corpus more amenable to learning, and then performing continued pretraining on the synthesized corpus.We instantiate this proposal with EntiGraph, a synthetic data augmentation algorithm that extracts salient entities from the source corpus and then generates diverse text by drawing connections between those entities.Synthetic continued pretraining with EntiGraph enables a language model to answer questions and follow generic instructions related to the source documents without access to them.If the source documents are instead available at inference time, we show that the knowledge acquired through our approach compounds with retrieval-augmented generation.To better understand these results, we build a simple mathematical model of EntiGraph, and show how synthetic data augmentation can "rearrange" knowledge to enable more data-efficient learning.

3694. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery

链接：https://iclr.cc/virtual/2025/poster/32108 abstract： The advancements of language language models (LLMs) have piqued growing interest in developing LLM-based language agents to automate scientific discovery end-to-end, which has sparked both excitement and skepticism about the true capabilities of such agents. In this work, we argue that for an agent to fully automate scientific discovery, it must be able to complete all essential tasks in the workflow. Thus, we call for rigorous assessment of agents on individual tasks in a scientific workflow before making bold claims on end-to-end automation. To this end, we present ScienceAgentBench, a new benchmark for evaluating language agents for data-driven scientific discovery. To ensure the

scientific authenticity and real-world relevance of our benchmark, we extract 102 tasks from 44 peer-reviewed publications in four disciplines and engage nine subject matter experts to validate them. We unify the target output for every task to a self-contained Python program file and employ an array of evaluation metrics to examine the generated programs, execution results, and costs. Each task goes through multiple rounds of manual validation by annotators and subject matter experts to ensure its annotation quality and scientific plausibility. We also propose two effective strategies to mitigate data contamination concerns. Using our benchmark, we evaluate five open-weight and proprietary LLMs, each with three frameworks: direct prompting, OpenHands, and self-debug. Given three attempts for each task, the best-performing agent can only solve 32.4% of the tasks independently and 34.3% with expert-provided knowledge. These results underscore the limited capacities of current language agents in generating code for data-driven discovery, let alone end-to-end automation for scientific research.

3695. Kronecker Mask and Interpretive Prompts are Language-Action Video Learners

链接：https://iclr.cc/virtual/2025/poster/29648 abstract： Contrastive language-image pretraining (CLIP) has significantly advanced image-based vision learning. A pressing topic subsequently arises: how can we effectively adapt CLIP to the video domain? Recent studies have focused on adjusting either the textual or visual branch of CLIP for action recognition. However, we argue that adaptations of both branches are crucial. In this paper, we propose a Contrastive Language-Action Video Learner (CLAVER), designed to shift CLIP's focus from the alignment of static visual objects and concrete nouns to the alignment of dynamic action behaviors and abstract verbs. Specifically, we introduce a novel Kronecker mask attention for temporal modeling. Our tailored Kronecker mask offers three benefits 1) it expands the temporal receptive field for each token, 2) it serves as an effective spatiotemporal heterogeneity inductive bias, mitigating the issue of spatiotemporal homogenization, and 3) it can be seamlessly plugged into transformer-based models. Regarding the textual branch, we leverage large language models to generate diverse, sentence-level and semantically rich interpretive prompts of actions, which shift the model's focus towards the verb comprehension. Extensive experiments on various benchmarks and learning scenarios demonstrate the superiority and generality of our approach. The code will be available soon.

3696. Speculative RAG: Enhancing Retrieval Augmented Generation through Drafting

链接：https://iclr.cc/virtual/2025/poster/27773 abstract： Retrieval augmented generation (RAG) combines the generative abilities of large language models (LLMs) with external knowledge sources to provide more accurate and up-to-date responses. Recent RAG advancements focus on improving retrieval outcomes through iterative LLM refinement or self-critique capabilities acquired through additional instruction tuning of LLMs. In this work, we introduce Speculative RAG - a framework that leverages a larger generalist LM to efficiently verify multiple RAG drafts produced in parallel by a smaller, distilled specialist LM. Each draft is generated from a distinct subset of retrieved documents, offering diverse perspectives on the evidence while reducing input token counts per draft. This approach enhances comprehension of each subset and mitigates potential position bias over long context. Our method accelerates RAG by delegating drafting to the smaller specialist LM, with the larger generalist LM performing a single verification pass over the drafts. Extensive experiments demonstrate that Speculative RAG achieves state-of-the-art performance with reduced latency on TriviaQA, MuSiQue, PopQA, PubHealth, and ARC-Challenge benchmarks. It notably enhances accuracy by up to 12.97% while reducing latency by 50.83% compared to conventional RAG systems on PubHealth.

3697. Rethinking Light Decoder-based Solvers for Vehicle Routing Problems

链接：https://iclr.cc/virtual/2025/poster/30992 abstract： Light decoder-based solvers have gained popularity for solving vehicle routing problems (VRPs) due to their efficiency and ease of integration with reinforcement learning algorithms. However, they often struggle with generalization to larger problem instances or different VRP variants. This paper revisits light decoder-based approaches, analyzing the implications of their reliance on static embeddings and the inherent challenges that arise. Specifically, we demonstrate that in the light decoder paradigm, the encoder is implicitly tasked with capturing information for all potential decision scenarios during solution construction within a single set of embeddings, resulting in high information density. Furthermore, our empirical analysis reveals that the overly simplistic decoder struggles to effectively utilize this dense information, particularly as task complexity increases, which limits generalization to out-of-distribution (OOD) settings. Building on these insights, we show that enhancing the decoder capacity, with a simple addition of identity mapping and a feed-forward layer, can considerably alleviate the generalization issue. Experimentally, our method significantly enhances the OOD generalization of light decoder-based approaches on large-scale instances and complex VRP variants, narrowing the gap with the heavy decoder paradigm. Our code is available at: https://github.com/ziweileonhuang/reld-nco.

3698. Relaxed Recursive Transformers: Effective Parameter Sharing with Layer-wise LoRA

链接：https://iclr.cc/virtual/2025/poster/29334 abstract： Large language models (LLMs) are expensive to deploy. Parameter sharing offers a possible path towards reducing their size and cost, but its effectiveness in modern LLMs remains fairly limited. In this work, we revisit "layer tying" as form of parameter sharing in Transformers, and introduce novel methods for converting existing LLMs into smaller "Recursive Transformers" that share parameters across layers, with minimal loss of performance.

Here, our Recursive Transformers are efficiently initialized from standard pretrained Transformers, but only use a single block of unique layers that is then repeated multiple times in a loop. We further improve performance by introducing Relaxed Recursive Transformers that add flexibility to the layer tying constraint via depth-wise low-rank adaptation (LoRA) modules, yet still preserve the compactness of the overall model. We show that our recursive models (e.g., recursive Gemma 1B) outperform both similar-sized vanilla pretrained models (such as TinyLlama 1.1B and Pythia 1B) and knowledge distillation baselines---and can even recover most of the performance of the original "full-size" model (e.g., Gemma 2B with no shared parameters). Finally, we propose Continuous Depth-wise Batching, a promising new inference paradigm enabled by the Recursive Transformer when paired with early exiting. In a theoretical analysis, we show that this has the potential to lead to significant (2-3$\times$) gains in inference throughput.

3699. ThinkBot: Embodied Instruction Following with Thought Chain Reasoning

链接：https://iclr.cc/virtual/2025/poster/28066 abstract： Embodied Instruction Following (EIF) requires agents to complete human instruction by interacting objects in complicated surrounding environments. Conventional methods directly consider the sparse human instruction to generate action plans for agents, which usually fail to achieve human goals because of the instruction incoherence in action descriptions. On the contrary, we propose ThinkBot that reasons the thought chain in human instruction to recover the missing action descriptions, so that the agent can successfully complete human goals by following the coherent instruction. Specifically, we first design an instruction completer based on large language models to recover the missing actions with interacted objects between consecutive human instruction, where the perceived surrounding environments and the completed sub-goals are considered for instruction completion. Based on the partially observed scene semantic maps, we present an object localizer to infer the position of interacted objects and the related Bayesian uncertainty for close-loop planning. Extensive experiments in the simulated environment show that our ThinkBot outperforms the state-of-the-art EIF methods by a sizable margin in both success rate and execution efficiency. Project page: https://guanxinglu.github.io/thinkbot/.

3700. EmbodiedSAM: Online Segment Any 3D Thing in Real Time

链接：https://iclr.cc/virtual/2025/poster/29314 abstract： Embodied tasks require the agent to fully understand 3D scenes simultaneously with its exploration, so an online, real-time, fine-grained and highly-generalized 3D perception model is desperately needed. Since high-quality 3D data is limited, directly training such a model in 3D is infeasible. Meanwhile, vision foundation models (VFM) has revolutionized the field of 2D computer vision with superior performance, which makes the use of VFM to assist embodied 3D perception a promising direction. However, most existing VFM-assisted 3D perception methods are either offline or too slow that cannot be applied in practical embodied tasks. In this paper, we aim to leverage Segment Anything Model (SAM) for real-time 3D instance segmentation in an online setting. This is a challenging problem since future frames are not available in the input streaming RGB-D video, and an instance may be observed in several frames so efficient object matching between frames is required. To address these challenges, we first propose a geometric-aware query lifting module to represent the 2D masks generated by SAM by 3D-aware queries, which is then iteratively refined by a dual-level query decoder. In this way, the 2D masks are transferred to fine-grained shapes on 3D point clouds. Benefit from the query representation for 3D masks, we can compute the similarity matrix between the 3D masks from different views by efficient matrix operation, which enables real-time inference. Experiments on ScanNet, ScanNet200, SceneNN and 3RScan show our method achieves state-of-the-art performance among online 3D perception models, even outperforming offline VFM-assisted 3D instance segmentation methods by a large margin. Our method also demonstrates great generalization ability in several zero-shot dataset transferring experiments and show great potential in data-efficient setting.

3701. Strong Preferences Affect the Robustness of Preference Models and Value Alignment

链接：https://iclr.cc/virtual/2025/poster/29453 abstract： Value alignment, which aims to ensure that large language models (LLMs) and other AI agents behave in accordance with human values, is critical for ensuring safety and trustworthiness of these systems. A key component of value alignment is the modeling of human preferences as a representation of human values. In this paper, we investigate the robustness of value alignment by examining the sensitivity of preference models. Specifically, we ask: how do changes in the probabilities of some preferences affect the predictions of these models for other preferences? To answer this question, we theoretically analyze the robustness of widely used preference models by examining their sensitivities to minor changes in preferences they model. Our findings reveal that, in the Bradley-Terry and the Placket-Luce model, the probability of a preference can change significantly as other preferences change, especially when these preferences are dominant (i.e., with probabilities near zero or one). We identify specific conditions where this sensitivity becomes significant for these models and discuss the practical implications for the robustness and safety of value alignment in AI systems.

3702. Breach By A Thousand Leaks: Unsafe Information Leakage in 'Safe' AI Responses

链接：https://iclr.cc/virtual/2025/poster/30768 abstract： Vulnerability of Frontier language models to misuse has prompted the development of safety measures like filters and alignment training seeking to ensure safety through robustness to adversarially crafted prompts. We assert that robustness is fundamentally insufficient for ensuring safety goals due to inferential threats from

dual-intent queries, with current defenses and evaluations failing to account for these risks. To quantify these risks, we introduce a new safety evaluation framework based on $\textit{impermissible information leakage}$ of model outputs and demonstrate how our proposed question-decomposition attack can extract dangerous knowledge from a censored LLM more effectively than traditional jailbreaking. Underlying our proposed evaluation method is a novel information-theoretic threat model of $\textit{inferential adversaries}$, distinguished from $\textit{security adversaries}$, such as jailbreaks, in that success involves inferring impermissible knowledge from victim outputs as opposed to forcing explicitly impermissible victim outputs. Through our information-theoretic framework, we show that ensuring safety against inferential adversaries requires defenses which bound impermissible information leakage, and, such defenses inevitably incur safety-utility trade-offs.

3703. Teaching LLMs How to Learn with Contextual Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/30340 abstract： Prompting Large Language Models (LLMs), or providing context on the expected model of operation, is an effective way to steer the outputs of such models to satisfy human desiderata after they have been trained. But in rapidly evolving domains, there is often need to fine-tune LLMs to improve either the kind of knowledge in their memory or their abilities to perform open ended reasoning in new domains. When human's learn new concepts, we often do so by linking the new material that we are studying to concepts we have already learned before. To that end, we ask, "can prompting help us teach LLMs how to learn". In this work, we study a novel generalization of instruction tuning, called contextual fine-tuning, to fine-tune LLMs. Our method leverages instructional prompts designed to mimic human cognitive strategies in learning and problem-solving to guide the learning process during training, aiming to improve the model’s interpretation and understanding of domain-specific knowledge. We empirically demonstrate that this simple yet effective modification improves the ability of LLMs to be fine-tuned rapidly on new datasets both within the medical and financial domains.

3704. Planning in Natural Language Improves LLM Search for Code Generation

链接：https://iclr.cc/virtual/2025/poster/31034 abstract： While scaling training compute has led to remarkable improvements in large language models (LLMs), scaling inference compute only recently began to yield analogous gains. We hypothesize that a core missing component is a lack of diverse LLM outputs, leading to inefficient search due to models repeatedly sampling highly similar, yet incorrect generations. We empirically demonstrate that this lack of diversity can be mitigated by searching over candidate plans for solving a problem in natural language. Based on this insight, we propose PlanSearch, a novel search algorithm which shows strong results across HumanEval+, MBPP+, and LiveCodeBench (a contamination-free benchmark for competitive coding). PlanSearch generates a diverse set of observations about the problem and uses these observations to construct plans for solving the problem. By searching over plans in natural language rather than directly over code solutions, PlanSearch explores a significantly more diverse range of potential solutions compared to baseline search methods. Using PlanSearch on top of Claude 3.5 Sonnet achieves a pass@200 of 77.0% on LiveCodeBench, outperforming both the best pass-rate achieved without any search (pass@1 = 41.4%) and using standard repeated sampling on top of existing non-search models (pass@200 = 60.6%). Finally, we show that, across all models, search algorithms, and benchmarks analyzed, we can accurately predict performance gains from search as a function of the diversity over generated ideas.

3705. BinaryDM: Accurate Weight Binarization for Efficient Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/29258 abstract： With the advancement of diffusion models (DMs) and the substantially increased computational requirements, quantization emerges as a practical solution to obtain compact and efficient low-bit DMs. However, the highly discrete representation leads to severe accuracy degradation, hindering the quantization of diffusion models to ultra-low bit-widths. This paper proposes a novel weight binarization approach for DMs, namely BinaryDM, pushing binarized DMs to be accurate and efficient by improving the representation and optimization. From the representation perspective, we present an Evolvable-Basis Binarizer (EBB) to enable a smooth evolution of DMs from full-precision to accurately binarized. EBB enhances information representation in the initial stage through the flexible combination of multiple binary bases and applies regularization to evolve into efficient single-basis binarization. The evolution only occurs in the head and tail of the DM architecture to retain the stability of training. From the optimization perspective, a Low-rank Representation Mimicking (LRM) is applied to assist the optimization of binarized DMs. The LRM mimics the representations of full-precision DMs in low-rank space, alleviating the direction ambiguity of the optimization process caused by fine-grained alignment. Comprehensive experiments demonstrate that BinaryDM achieves significant accuracy and efficiency gains compared to SOTA quantization methods of DMs under ultra-low bit-widths. With 1-bit weight and 4-bit activation (W1A4), BinaryDM achieves as low as 7.74 FID and saves the performance from collapse (baseline FID 10.87). As the first binarization method for diffusion models, W1A4 BinaryDM achieves impressive 15.2x OPs and 29.2x model size savings, showcasing its substantial potential for edge deployment.

3706. A Theoretical Analysis of Self-Supervised Learning for Vision Transformers

链接：https://iclr.cc/virtual/2025/poster/30614 abstract： Self-supervised learning has become a cornerstone in computer vision, primarily divided into reconstruction-based methods like masked autoencoders (MAE) and discriminative methods such as contrastive learning (CL). Recent empirical observations reveal that MAE and CL capture different types of representations: CL tends to focus on global patterns, while MAE adeptly captures both global and subtle local information simultaneously.

Despite a flurry of recent empirical investigations to shed light on this difference, theoretical understanding remains limited, especially on the dominant architecture vision transformers (ViTs). In this paper, to provide rigorous insights, we model the visual data distribution by considering two types of spatial features: dominant global features and comparatively minuscule local features, and study the impact of imbalance among these features. We analyze the training dynamics of one-layer softmax-based ViTs on both MAE and CL objectives using gradient descent. Our analysis shows that as the degree of feature imbalance varies, ViTs trained with the MAE objective effectively learn both global and local features to achieve near-optimal reconstruction, while the CL-trained ViTs favor predominantly global features, even under mild imbalance. These results provide a theoretical explanation for distinct behaviors of MAE and CL observed in empirical studies.

3707. Open-Vocabulary Customization from CLIP via Data-Free Knowledge Distillation

链接：https://iclr.cc/virtual/2025/poster/31193 abstract： Vision-language models such as CLIP have demonstrated strong zero-shot performance, but their considerable size and inefficient inference limit customizable deployment for users. While knowledge distillation is a solution, it still requires the original data, which is not always available due to copyrights and privacy concerns. For many users seeking open-vocabulary customization, Data-Free Knowledge Distillation (DFKD) emerges as a promising direction. Upon rethinking DFKD, we find that existing methods fail on CLIP due to their heavy reliance on BatchNorm layers, which are unexpectedly unusable in CLIP. Based on our findings, we adopt image-text matching to achieve DFKD for CLIP, enabling customization based on arbitrary class texts. This involves (i) inversing a surrogate dataset from CLIP based on text prompts; and (ii) distilling a student model from CLIP using the surrogate dataset. Specifically, we introduce style dictionary diversification to enhance the diversity of synthetic images. To prevent uncontrollable semantics introduced by diversification, we propose a class consistency maintaining strategy to ensure the consistency of synthetic images. Based on synthetic images with various styles, we further propose meta knowledge distillation to train the student model with good generalization ability. Moreover, we introduce a simple yet effective method to enable customization based on few example images. Comprehensive experiments showcase the superiority of our approach across twelve customized tasks, achieving a 9.33\% improvement compared to existing DFKD methods.

3708. Bayesian WeakS-to-Strong from Text Classification to Generation

链接：https://iclr.cc/virtual/2025/poster/28306 abstract： Advances in large language models raise the question of how alignment techniques will adapt as models become increasingly complex and humans will only be able to supervise them weakly. Weak-to-Strong mimics such a scenario where weak model supervision attempts to harness the full capabilities of a much stronger model. This work extends Weak-to-Strong to WeakS-to-Strong by exploring an ensemble of weak models which simulate the variability in human opinions. Confidence scores are estimated using a Bayesian approach to guide the WeakS-to-Strong generalization. Furthermore, we extend the application of WeakS-to-Strong from text classification tasks to text generation tasks where more advanced strategies are investigated for supervision. Moreover, direct preference optimization is applied to advance the student model's preference learning, beyond the basic learning framework of teacher forcing. Results demonstrate the effectiveness of the proposed approach for the reliability of a strong student model, showing potential for superalignment.

3709. Debiasing Federated Learning with Correlated Client Participation

链接：https://iclr.cc/virtual/2025/poster/30673 abstract： In cross-device federated learning (FL) with millions of mobile clients, only a small subset of clients participate in training in every communication round, and Federated Averaging (FedAvg) is the most popular algorithm in practice. Existing analyses of FedAvg usually assume the participating clients are independently sampled in each round from a uniform distribution, which does not reflect real-world scenarios. This paper introduces a theoretical framework that models client participation in FL as a Markov chain to study optimization convergence when clients have non-uniform and correlated participation across rounds. We apply this framework to analyze a more practical pattern: every client must wait a minimum number of R rounds (minimum separation) before re-participating. We theoretically prove and empirically observe that increasing minimum separation reduces the bias induced by intrinsic non-uniformity of client availability in cross-device FL systems. Furthermore, we develop an effective debiasing algorithm for FedAvg that provably converges to the unbiased optimal solution under arbitrary minimum separation and unknown client availability distribution.

3710. TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models

链接：https://iclr.cc/virtual/2025/poster/32069 abstract： Existing benchmarks often highlight the remarkable performance achieved by state-of-the-art Multimodal Foundation Models (MFMs) in leveraging temporal context for video understanding.However, how well do the models truly perform visual temporal reasoning?Our study of existing benchmarks shows that this capability of MFMs is likely overestimated as many questions can be solved by using a single, few, or out-of-order frames. To systematically examine current visual temporal reasoning tasks, we propose three principles with corresponding metrics:(1) Multi-Frame Gain,(2) Frame Order Sensitivity,and (3) Frame Information Disparity.Following these principles, we introduce TOMATO, TempOral Reasoning MultimodAl EvaluaTiOn, a novel benchmark crafted to rigorously assess MFMs' temporal reasoning capabilities in video understanding.TOMATO comprises 1,484 carefully curated, human-annotated questions spanning six tasks (i.e. action count, direction, rotation, shape & trend, velocity & frequency, and visual

cues), applied to 1,417 videos, including 805 self-recorded and -generated videos, that encompass human-centric, real-world, and simulated scenarios. Our comprehensive evaluation reveals a human-model performance gap of 57.3% with the best-performing model.Moreover, our in-depth analysis uncovers more fundamental limitations beyond this gap in current MFMs. While they can accurately recognize events in isolated frames, they fail to interpret these frames as a continuous sequence.We believe TOMATO will serve as a crucial testbed for evaluating the next-generation MFMs and as a call to the community to develop AI systems capable of comprehending the human world dynamics through the video modality.

3711. Towards counterfactual fairness through auxiliary variables

链接：https://iclr.cc/virtual/2025/poster/30258 abstract： The challenge of balancing fairness and predictive accuracy in machine learning models, especially when sensitive attributes such as race, gender, or age are considered, has motivated substantial research in recent years. Counterfactual fairness ensures that predictions remain consistent across counterfactual variations of sensitive attributes, which is a crucial concept in addressing societal biases. However, existing counterfactual fairness approaches usually overlook intrinsic information about sensitive features, limiting their ability to achieve fairness while simultaneously maintaining performance. To tackle this challenge, we introduce EXOgenous Causal reasoning (EXOC), a novel causal reasoning framework motivated by exogenous variables. It leverages auxiliary variables to uncover intrinsic properties that give rise to sensitive attributes. Our framework explicitly defines an auxiliary node and a control node that contribute to counterfactual fairness and control the information flow within the model. Our evaluation, conducted on synthetic and real-world datasets, validates EXOC's superiority, showing that it outperforms state-of-the-art approaches in achieving counterfactual fairness without sacrificing accuracy. Our code is available at https://github.com/CASE-Lab-UMD/counterfactualfairness2025.

3712. SG-I2V: Self-Guided Trajectory Control in Image-to-Video Generation

链接：https://iclr.cc/virtual/2025/poster/27979 abstract： Methods for image-to-video generation have achieved impressive, photo-realistic quality. However, adjusting specific elements in generated videos, such as object motion or camera movement, is often a tedious process of trial and error, e.g., involving re-generating videos with different random seeds. Recent techniques address this issue by fine-tuning a pre-trained model to follow conditioning signals, such as bounding boxes or point trajectories. Yet, this fine-tuning procedure can be computationally expensive, and it requires datasets with annotated object motion, which can be difficult to procure. In this work, we introduce SG-I2V, a framework for controllable image-to-video generation that is self-guided—offering zero-shot control by relying solely on the knowledge present in a pre-trained image-to-video diffusion model without the need for fine-tuning or external knowledge. Our zero-shot method outperforms unsupervised baselines while significantly narrowing down the performance gap with supervised models in terms of visual quality and motion fidelity.Additional details and video results are available on our project page: https://kmcode1.github.io/Projects/SG-I2V

3713. WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling

链接：https://iclr.cc/virtual/2025/poster/27745 abstract： Language models have been effectively applied to modeling natural signals, such as images, video, speech, and audio. A crucial component of these models is the codec tokenizer, which compresses high-dimensional natural signals into lower-dimensional discrete tokens. In this paper, we introduce WavTokenizer, which offers several advantages over previous SOTA acoustic codec models in the audio domain: 1) extreme compression. By compressing the layers of quantizers and the temporal dimension of the discrete codec, one-second audio of 24kHz sampling rate requires only a single quantizer with 40 or 75 tokens. 2) improved subjective quality. Despite the reduced number of tokens, WavTokenizer achieves state-of-the-art reconstruction quality with outstanding UTMOS scores and inherently contains richer semantic information. Specifically, we achieve these results by designing a broader VQ space, extended contextual windows, and improved attention networks, as well as introducing a powerful multi-scale discriminator and an inverse Fourier transform structure. We conducted extensive reconstruction experiments in the domains of speech, audio, and music. WavTokenizer exhibited strong performance across various objective and subjective metrics compared to state-of-the-art models. We also tested semantic information, VQ utilization, and adaptability to generative models. Comprehensive ablation studies confirm the necessity of each module in WavTokenizer. The code is available at https://github.com/jishengpeng/WavTokenizer.

3714. Improving Neural Optimal Transport via Displacement Interpolation

链接：https://iclr.cc/virtual/2025/poster/30504 abstract： Optimal Transport (OT) theory investigates the cost-minimizing transport map that moves a source distribution to a target distribution. Recently, several approaches have emerged for learning the optimal transport map for a given cost function using neural networks. We refer to these approaches as the OT Map. OT Map provides a powerful tool for diverse machine learning tasks, such as generative modeling and unpaired image-to-image translation. However, existing methods that utilize max-min optimization often experience training instability and sensitivity to hyperparameters. In this paper, we propose a novel method to improve stability and achieve a better approximation of the OT Map by exploiting displacement interpolation, dubbed Displacement Interpolation Optimal Transport Model (DIOTM). We derive the dual formulation of displacement interpolation at specific time t and prove how these dual problems are related across time. This result allows us to utilize the entire trajectory of displacement interpolation in learning the OT Map. Our method improves the training stability and achieves superior results in estimating optimal transport maps. We demonstrate that DIOTM outperforms existing OT-based models on image-to-image translation tasks.

3715. Towards Continuous Reuse of Graph Models via Holistic Memory Diversification

链接：https://iclr.cc/virtual/2025/poster/29744 abstract： This paper addresses the challenge of incremental learning in growing graphs with increasingly complex tasks. The goal is to continuously train a graph model to handle new tasks while retaining proficiency in previous tasks via memory replay. Existing methods usually overlook the importance of memory diversity, limiting in selecting high-quality memory from previous tasks and remembering broad previous knowledge within the scarce memory on graphs. To address that, we introduce a novel holistic Diversified Memory Selection and Generation (DMSG) framework for incremental learning in graphs, which first introduces a buffer selection strategy that considers both intra-class and inter-class diversities, employing an efficient greedy algorithm for sampling representative training nodes from graphs into memory buffers after learning each new task. Then, to adequately rememorize the knowledge preserved in the memory buffer when learning new tasks, a diversified memory generation replay method is introduced. This method utilizes a variational layer to generate the distribution of buffer node embeddings and sample synthesized ones for replaying. Furthermore, an adversarial variational embedding learning method and a reconstruction-based decoder are proposed to maintain the integrity and consolidate the generalization of the synthesized node embeddings, respectively. Extensive experimental results on publicly accessible datasets demonstrate the superiority of DMSG over state-of-the-art methods.

3716. SleepSMC: Ubiquitous Sleep Staging via Supervised Multimodal Coordination

链接：https://iclr.cc/virtual/2025/poster/30593 abstract： Sleep staging is critical for assessing sleep quality and tracking health. Polysomnography (PSG) provides comprehensive multimodal sleep-related information, but its complexity and impracticality limit its practical use in daily and ubiquitous monitoring. Conversely, unimodal devices offer more convenience but less accuracy. Existing multimodal learning paradigms typically assume that the data types remain consistent between the training and testing phases. This makes it challenging to leverage information from other modalities in ubiquitous scenarios (e.g., at home) where only one modality is available. To address this issue, we introduce a novel framework for ubiquitous Sleep staging via Supervised Multimodal Coordination, called SleepSMC. To capture category-related consistency and complementarity across modality-level instances, we propose supervised modality-level instance contrastive coordination. Specifically, modality-level instances within the same category are considered positive pairs, while those from different categories are considered negative pairs. To explore the varying reliability of auxiliary modalities, we calculate uncertainty estimates based on the variance in confidence scores for correct predictions during multiple rounds of random masks. These uncertainty estimates are employed to assign adaptive weights to multiple auxiliary modalities during contrastive learning, ensuring that the primary modality learns from high-quality, category-related features. Experimental results on four public datasets, ISRUC-S3, MASS-SS3, Sleep-EDF-78, and ISRUC-S1, show that SleepSMC achieves state-of-the-art cross-subject performance. SleepSMC significantly improves performance when only one modality is present during testing, making it suitable for ubiquitous sleep monitoring.

3717. EqNIO: Subequivariant Neural Inertial Odometry

链接：https://iclr.cc/virtual/2025/poster/30531 abstract： Neural network-based odometry using accelerometer and gyroscope readings from a single IMU can achieve robust, and low-drift localization capabilities, through the use of *neural displacement priors (NDPs)*. These priors learn to produce denoised displacement measurements but need to ignore data variations due to specific IMU mount orientation and motion directions, hindering generalization.This work introduces EqNIO, which addresses this challenge with *canonical displacement priors*, i.e., priors that are invariant to the orientation of the gravity-aligned frame in which the IMU data is expressed. We train such priors on IMU measurements, that are mapped into a learnable canonical frame, which is uniquely defined via three axes: the first is gravity, making the frame gravity aligned, while the second and third are predicted from IMU data. The outputs (displacement and covariance) are mapped back to the original gravity-aligned frame. To maximize generalization, we find that these learnable frames must transform equivariantly with global gravity-preserving roto-reflections from the subgroup $O_g(3)\subset O(3)$, acting on the trajectory, rendering the NDP $O(3)$-*subequivariant*. We tailor specific linear, convolutional, and non-linear layers that commute with the actions of the group. Moreover, we introduce a bijective decomposition of angular rates into vectors that transform similarly to accelerations, allowing us to leverage both measurement types. Natively, angular rates would need to be inverted upon reflection, unlike acceleration, which hinders their joint processing. We highlight EqNIO's flexibility and generalization capabilities by applying it to both filter-based (TLIO), and end-to-end (RONIN) architectures, and outperforming existing methods that use _soft equivariance from auxiliary losses or data augmentation on various datasets. We believe this work paves the way for low-drift and generalizable neural inertial odometry on edge devices. The project details and code can be found at https://github.com/RoyinaJayanth/EqNIO.

3718. OmniRe: Omni Urban Scene Reconstruction

链接：https://iclr.cc/virtual/2025/poster/31225 abstract： We introduce OmniRe, a comprehensive system for efficiently creating high-fidelity digital twins of dynamic real-world scenes from on-device logs. Recent methods using neural fields or Gaussian Splatting primarily focus on vehicles, hindering a holistic framework for all dynamic foregrounds demanded by downstream applications, e.g., the simulation of human behavior. OmniRe extends beyond vehicle modeling to enable accurate, full-length reconstruction of diverse dynamic objects in urban scenes. Our approach builds scene graphs on 3DGS and constructs multiple Gaussian representations in canonical spaces that model various dynamic actors, including vehicles, pedestrians, cyclists, and others. OmniRe allows holistically reconstructing any dynamic object in the scene, enabling advanced simulations (~60 Hz) that

include human-participated scenarios, such as pedestrian behavior simulation and human-vehicle interaction. This comprehensive simulation capability is unmatched by existing methods. Extensive evaluations on the Waymo dataset show that our approach outperforms prior state-of-the-art methods quantitatively and qualitatively by a large margin. We further extend our results to 5 additional popular driving datasets to demonstrate its generalizability on common urban scenes. Code and results are available at omnire.

3719. EG4D: Explicit Generation of 4D Object without Score Distillation

链接：https://iclr.cc/virtual/2025/poster/27956 abstract： In recent years, the increasing demand for dynamic 3D assets in design and gaming applications has given rise to powerful generative pipelines capable of synthesizing high-quality 4D objects. Previous methods generally rely on score distillation sampling (SDS) algorithm to infer the unseen views and motion of 4D objects, thus leading to unsatisfactory results with defects like over-saturation and Janus problem. Therefore, inspired by recent progress of video diffusion models, we propose to optimize a 4D representation by explicitly generating multi-view videos from one input image. However, it is far from trivial to handle practical challenges faced by such a pipeline, including dramatic temporal inconsistency, inter-frame geometry and texture diversity, and semantic defects brought by video generation results. To address these issues, we propose EG4D, a novel multi-stage framework that generates high-quality and consistent 4D assets without score distillation. Specifically, collaborative techniques and solutions are developed, including an attention injection strategy to synthesize temporal-consistent multi-view videos, a robust and efficient dynamic reconstruction method based on Gaussian Splatting, and a refinement stage with diffusion prior for semantic restoration. The qualitative comparisons and quantitative results demonstrate that our framework outperforms the baselines in generation quality by a considerable margin.

3720. Unsupervised Disentanglement of Content and Style via Variance-Invariance Constraints

链接：https://iclr.cc/virtual/2025/poster/29968 abstract： We contribute an unsupervised method that effectively learns disentangled content and style representations from sequences of observations. Unlike most disentanglement algorithms that rely on domain-specific labels or knowledge, our method is based on the insight of domain-general statistical differences between content and style --- content varies more among different fragments within a sample but maintains an invariant vocabulary across data samples, whereas style remains relatively invariant within a sample but exhibits more significant variation across different samples. We integrate such inductive bias into an encoder-decoder architecture and name our method after V3 (variance-versus-invariance). Experimental results show that V3 generalizes across multiple domains and modalities, successfully learning disentangled content and style representations, such as pitch and timbre from music audio, digit and color from images of hand-written digits, and action and character appearance from simple animations. V3 demonstrates strong disentanglement performance compared to existing unsupervised methods, along with superior out-of-distribution generalization and few-shot learning capabilities compared to supervised counterparts. Lastly, symbolic-level interpretability emerges in the learned content codebook, forging a near one-to-one alignment between machine representation and human knowledge.

3721. DOTS: Learning to Reason Dynamically in LLMs via Optimal Reasoning Trajectories Search

链接：https://iclr.cc/virtual/2025/poster/28026 abstract： Enhancing the capability of large language models (LLMs) in reasoning has gained significant attention in recent years. Previous studies have demonstrated the effectiveness of various prompting strategies in aiding LLMs in reasoning (called "reasoning actions"), such as step-by-step thinking, reflecting before answering, solving with programs, and their combinations. However, these approaches often applied static, predefined reasoning actions uniformly to all questions, without considering the specific characteristics of each question or the capability of the task-solving LLM. In this paper, we propose DOTS, an approach enabling LLMs to reason Dynamically via Optimal reasoning Trajectories Search, tailored to the specific characteristics of each question and the inherent capability of the task-solving LLM. Our approach involves three key steps: i) defining atomic reasoning action modules that can be composed into various reasoning action trajectories; ii) searching for the optimal action trajectory for each training question through iterative exploration and evaluation for the specific task-solving LLM; and iii) using the collected optimal trajectories to train an LLM to plan for the reasoning trajectories of unseen questions. In particular, we propose two learning paradigms, i.e., fine-tuning an external LLM as a planner to guide the task-solving LLM, or directly fine-tuning the task-solving LLM with an internalized capability for reasoning actions planning. Our experiments across eight reasoning tasks show that our method consistently outperforms static reasoning techniques and the vanilla instruction tuning approach. Further analysis reveals that our method enables LLMs to adjust their computation based on problem complexity, allocating deeper thinking and reasoning to harder problems.

3722. Cauchy-Schwarz Regularizers

链接：https://iclr.cc/virtual/2025/poster/30046 abstract： We introduce a novel class of regularization functions, called Cauchy–Schwarz (CS) regularizers, which can be designed to induce a wide range of properties in solution vectors of optimization problems. To demonstrate the versatility of CS regularizers, we derive regularization functions that promote discrete-valued vectors, eigenvectors of a given matrix, and orthogonal matrices. The resulting CS regularizers are simple, differentiable, and can be free of spurious stationary points, making them suitable for gradient-based solvers and large-scale optimization problems. In addition, CS regularizers automatically adapt to the appropriate scale, which is, for example, beneficial when

discretizing the weights of neural networks. To demonstrate the efficacy of CS regularizers, we provide results for solving underdetermined systems of linear equations and weight quantization in neural networks. Furthermore, we discuss specializations, variations, and generalizations, which lead to an even broader class of new and possibly more powerful regularizers.

3723. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28330 abstract： Multi-agent reinforcement learning (MARL) methods struggle with the non-stationarity of multi-agent systems and fail to adaptively learn online when tested with novel agents. Here, we leverage large language models (LLMs) to create an autonomous agent that can handle these challenges. Our agent, Hypothetical Minds, consists of a cognitively-inspired architecture, featuring modular components for perception, memory, and hierarchical planning over two levels of abstraction. We introduce the Theory of Mind module that scaffolds the high-level planning process by generating hypotheses about other agents' strategies in natural language. It then evaluates and iteratively refines these hypotheses by reinforcing hypotheses that make correct predictions about the other agents' behavior. Hypothetical Minds significantly improves performance over previous LLM-agent and RL baselines on a range of competitive, mixed motive, and collaborative domains in the Melting Pot benchmark, including both dyadic and population-based environments. Additionally, comparisons against LLM-agent baselines and ablations reveal the importance of hypothesis evaluation and refinement for succeeding on complex scenarios.

3724. T-Stitch: Accelerating Sampling in Pre-Trained Diffusion Models with Trajectory Stitching

链接：https://iclr.cc/virtual/2025/poster/31118 abstract： Sampling from diffusion probabilistic models (DPMs) is often expensive for high-quality image generation and typically requires many steps with a large model. In this paper, we introduce sampling Trajectory Stitching (T-Stitch), a simple yet efficient technique to improve the sampling efficiency with little or no generation degradation. Instead of solely using a large DPM for the entire sampling trajectory, T-Stitch first leverages a smaller DPM in the initial steps as a cheap drop-in replacement of the larger DPM and switches to the larger DPM at a later stage. Our key insight is that different diffusion models learn similar encodings under the same training data distribution and smaller models are capable of generating good global structures in the early steps. Extensive experiments demonstrate that T-Stitch is training-free, generally applicable for different architectures, and complements most existing fast sampling techniques with flexible speed and quality trade-offs. On DiT-XL, for example, 40% of the early timesteps can be safely replaced with a 10x faster DiT-S without performance drop on class-conditional ImageNet generation. We further show that our method can also be used as a drop-in technique to not only accelerate the popular pretrained stable diffusion (SD) models but also improve the prompt alignment of stylized SD models from the public model zoo. Finally, the explicit model allocation strategy of T-Stitch significantly reduces the need of training or searching, delivering high deployment efficiency.

3725. PaRa: Personalizing Text-to-Image Diffusion via Parameter Rank Reduction

链接：https://iclr.cc/virtual/2025/poster/30047 abstract： Personalizing a large-scale pretrained Text-to-Image (T2I) diffusion model is chal-lenging as it typically struggles to make an appropriate trade-off between its trainingdata distribution and the target distribution, i.e., learning a novel concept with only afew target images to achieve personalization (aligning with the personalized target)while preserving text editability (aligning with diverse text prompts). In this paper,we propose PaRa, an effective and efficient Parameter Rank Reduction approachfor T2I model personalization by explicitly controlling the rank of the diffusionmodel parameters to restrict its initial diverse generation space into a small andwell-balanced target space. Our design is motivated by the fact that taming a T2Imodel toward a novel concept such as a specific art style implies a small generationspace. To this end, by reducing the rank of model parameters during finetuning, wecan effectively constrain the space of the denoising sampling trajectories towardsthe target. With comprehensive experiments, we show that PaRa achieves greatadvantages over existing finetuning approaches on single/multi-subject generationas well as single-image editing. Notably, compared to the prevailing fine-tuningtechnique LoRA, PaRa achieves better parameter efficiency (2× fewer learnableparameters) and much better target image alignment.

3726. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28277 abstract： Multi-modal Large Language Models have demonstrated remarkable capabilities in executing instructions for a variety of single-image tasks. Despite this progress, significant challenges remain in modeling long image sequences. In this work, we introduce the versatile multi-modal large language model, mPLUG-Owl3, which enhances the capability for long image-sequence understanding in scenarios that incorporate retrieved image-text knowledge, multimodal in-context examples, and lengthy videos. Specifically, we propose novel hyper attention blocks to efficiently integrate vision and language into a common language-guided semantic space, thereby facilitating the processing of extended multi-image scenarios. We conduct evaluations on 21 benchmarks that cover single/multi-image, and short/long video understanding. mPLUG-Owl3 achieves competitive performance with the state-of-the-art methods while reducing inference time

and memory usage by 87.8\% and 48.5\% in average. Moreover, we propose a Distractor Resistance evaluation to assess the ability of models to maintain focus amidst distractions. mPLUG-Owl3 also demonstrates outstanding performance in distractor resistance on ultra-long visual sequence inputs. We hope that mPLUG-Owl3 can contribute to the development of more efficient and powerful multimodal large language models.

3727. INS: Interaction-aware Synthesis to Enhance Offline Multi-agent Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28551 abstract： Data scarcity in offline multi-agent reinforcement learning (MARL) is a key challenge for real-world applications. Recent advances in offline single-agent reinforcement learning (RL) demonstrate the potential of data synthesis to mitigate this issue.However, in multi-agent systems, interactions between agents introduce additional challenges. These interactions complicate the synthesis of multi-agent datasets, leading to data distortion when inter-agent interactions are neglected. Furthermore, the quality of the synthetic dataset is often constrained by the original dataset. To address these challenges, we propose INteraction-aware Synthesis (INS), which synthesizes high-quality multi-agent datasets using diffusion models. Recognizing the sparsity of inter-agent interactions, INS employs a sparse attention mechanism to capture these interactions, ensuring that the synthetic dataset reflects the underlying agent dynamics. To overcome the limitation of diffusion models requiring continuous variables, INS implements a bit action module, enabling compatibility with both discrete and continuous action spaces. Additionally, we incorporate a select mechanism to prioritize transitions with higher estimated values, further enhancing the dataset quality. Experimental results across multiple datasets in MPE and SMAC environments demonstrate that INS consistently outperforms existing methods, resulting in improved downstream policy performance and superior dataset metrics. Notably, INS can synthesize high-quality data using only 10% of the original dataset, highlighting its efficiency in data-limited scenarios.

3728. A Periodic Bayesian Flow for Material Generation

链接：https://iclr.cc/virtual/2025/poster/29962 abstract： Generative modeling of crystal data distribution is an important yet challenging task due to the unique periodic physical symmetry of crystals. Diffusion-based methods have shown early promise in modeling crystal distribution. More recently, Bayesian Flow Networks were introduced to aggregate noisy latent variables, resulting in a variance-reduced parameter space that has been shown to be advantageous for modeling Euclidean data distributions with structural constraints (Song, et al.,2023). Inspired by this, we seek to unlock its potential for modeling variables located in non-Euclidean manifolds e.g. those within crystal structures, by overcoming challenging theoretical issues. We introduce CrysBFN, a novel crystal generation method by proposing a periodic Bayesian flow, which essentially differs from the original Gaussian-based BFN by exhibiting non-monotonic entropy dynamics. To successfully realize the concept of periodic Bayesian flow, CrysBFN integrates a new entropy conditioning mechanism and empirically demonstrates its significance compared to time-conditioning. Extensive experiments over both crystal ab initio generation and crystal structure prediction tasks demonstrate the superiority of CrysBFN, which consistently achieves new state-of-the-art on all benchmarks. Surprisingly, we found that CrysBFN enjoys a significant improvement in sampling efficiency, e.g., 200x speedup (10 v.s. 2000 steps network forwards) compared with previous Diffusion-based methods on MP-20 dataset.

3729. MaskGCT: Zero-Shot Text-to-Speech with Masked Generative Codec Transformer

链接：https://iclr.cc/virtual/2025/poster/30376 abstract：

3730. SWIFT: On-the-Fly Self-Speculative Decoding for LLM Inference Acceleration

链接：https://iclr.cc/virtual/2025/poster/30412 abstract： Speculative decoding (SD) has emerged as a widely used paradigm to accelerate LLM inference without compromising quality. It works by first employing a compact model to draft multiple tokens efficiently and then using the target LLM to verify them in parallel. While this technique has achieved notable speedups, most existing approaches necessitate either additional parameters or extensive training to construct effective draft models, thereby restricting their applicability across different LLMs and tasks. To address this limitation, we explore a novel plug-and-play SD solution with layer-skipping, which skips intermediate layers of the target LLM as the compact draft model. Our analysis reveals that LLMs exhibit great potential for self-acceleration through layer sparsity and the task-specific nature of this sparsity. Building on these insights, we introduce SWIFT, an on-the-fly self-speculative decoding algorithm that adaptively selects intermediate layers of LLMs to skip during inference. SWIFT does not require auxiliary models or additional training, making it a plug-and-play solution for accelerating LLM inference across diverse input data streams. Our extensive experiments across a wide range of models and downstream tasks demonstrate that SWIFT can achieve over a $1.3\times$\sim$1.6\times$ speedup while preserving the original distribution of the generated text. We release our code in https://github.com/hemingkx/SWIFT.

3731. FlexCAD: Unified and Versatile Controllable CAD Generation with Fine-tuned Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29237 abstract： Recently, there is a growing interest in creating computer-aided design (CAD) models based on user intent, known as controllable CAD generation. Existing work offers limited controllability and needs separate models for different types of control, reducing efficiency and practicality. To achieve controllable generation across all CAD construction hierarchies, such as sketch-extrusion, extrusion, sketch, face, loop and curve, we propose FlexCAD, a unified model by fine-tuning large language models (LLMs). First, to enhance comprehension by LLMs, we represent a CAD model as a structured text by abstracting each hierarchy as a sequence of text tokens. Second, to address various controllable generation tasks in a unified model, we introduce a hierarchy-aware masking strategy. Specifically, during training, we mask a hierarchy-aware field in the CAD text with a mask token. This field, composed of a sequence of tokens, can be set flexibly to represent various hierarchies. Subsequently, we ask LLMs to predict this masked field. During inference, the user intent is converted into a CAD text with a mask token replacing the part the user wants to modify, which is then fed into FlexCAD to generate new CAD models. Comprehensive experiments on public dataset demonstrate the effectiveness of FlexCAD in both generation quality and controllability.

3732. Don't Take Things Out of Context: Attention Intervention for Enhancing Chain-of-Thought Reasoning in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29384 abstract： Few-shot Chain-of-Thought (CoT) significantly enhances the reasoning capabilities of large language models (LLMs), functioning as a whole to guide these models in generating reasoning steps toward final answers. However, we observe that isolated segments, words, or tokens within CoT demonstrations can unexpectedly disrupt the generation process of LLMs. The model may overly concentrate on certain local information present in the demonstration, introducing irrelevant noise into the reasoning process and potentially leading to incorrect answers. In this paper, we investigate the underlying mechanism of CoT through dynamically tracing and manipulating the inner workings of LLMs at each output step, which demonstrates that tokens exhibiting specific attention characteristics are more likely to induce the model to take things out of context; these tokens directly attend to the hidden states tied with prediction, without substantial integration of non-local information. Building upon these insights, we propose a Few-shot Attention Intervention method (FAI) that dynamically analyzes the attention patterns of demonstrations to accurately identify these tokens and subsequently make targeted adjustments to the attention weights to effectively suppress their distracting effect on LLMs. Comprehensive experiments across multiple benchmarks demonstrate consistent improvements over baseline methods, with a remarkable 5.91\% improvement on the AQuA dataset, further highlighting the effectiveness of FAI.

3733. Improving Complex Reasoning with Dynamic Prompt Corruption: A Soft Prompt Optimization Approach

链接：https://iclr.cc/virtual/2025/poster/28777 abstract： Prompt Tuning (PT) has emerged as a promising Parameter-Efficient Fine-Tuning (PEFT) approach by appending trainable continuous prompt vectors to the input, maintaining competitive performance with significantly fewer trainable parameters. While PT has shown effectiveness in enhancing task performance, particularly for classification tasks, its application to complex reasoning tasks has been largely overlooked. Our investigation reveals that PT provides limited improvement and may even degrade performance in reasoning tasks. This phenomenon suggests that soft prompts can positively impact certain instances while negatively affecting others, particularly during the latter stages of reasoning.To address these challenges, we propose a novel method called Dynamic Prompt Corruption (DPC), which seeks to optimize the use of soft prompts in reasoning tasks. DPC dynamically adjusts the influence of soft prompts based on their impact on the reasoning process. Specifically, it involves two key components: Dynamic Trigger and Dynamic Corruption. Dynamic Trigger measures the influence of soft prompts, determining whether their impact is beneficial or detrimental. Dynamic Corruption mitigates the negative effects of soft prompts by selectively masking key tokens that interfere with the reasoning process.We validate our approach through extensive experiments on various large language models (LLMs) and reasoning tasks, including GSM8K, MATH, and AQuA. The results demonstrate that Dynamic Prompt Corruption consistently improves the performance of LLMs, achieving 4\%-8\% accuracy gains compared to standard prompt tuning. These findings highlight the effectiveness of our approach and its potential to enhance complex reasoning in LLMs.

3734. Constraint-Conditioned Actor-Critic for Offline Safe Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28392 abstract： Offline safe reinforcement learning (OSRL) aims to learn policies with high rewards while satisfying safety constraints solely from data collected offline. However, the learned policies often struggle to handle states and actions that are not present or out-of-distribution (OOD) from the offline dataset, which can result in violation of the safety constraints or overly conservative behaviors during their online deployment. Moreover, many existing methods are unable to learn policies that can adapt to varying constraint thresholds. To address these challenges, we propose constraint-conditioned actor-critic (CCAC), a novel OSRL method that models the relationship between state-action distributions and safety constraints, and leverages this relationship to regularize critics and policy learning. CCAC learns policies that can effectively handle OOD data and adapt to varying constraint thresholds. Empirical evaluations on the $\texttt{DSRL}$ benchmarks show that CCAC significantly outperforms existing methods for learning adaptive, safe, and high-reward policies.

3735. Language Model Alignment in Multilingual Trolley Problems

链接：https://iclr.cc/virtual/2025/poster/32081 abstract： We evaluate the moral alignment of large language models (LLMs) with human preferences in multilingual trolley problems. Building on the Moral Machine experiment, which captures over 40 million human judgments across 200+ countries, we develop a cross-lingual corpus of moral dilemma vignettes in over 100 languages called MultiTP. This dataset enables the assessment of LLMs' decision-making processes in diverse linguistic contexts. Our analysis explores the alignment of 19 different LLMs with human judgments, capturing preferences across six moral dimensions: species, gender, fitness, status, age, and the number of lives involved. By correlating these preferences with the demographic distribution of language speakers and examining the consistency of LLM responses to various prompt paraphrasings, our findings provide insights into cross-lingual and ethical biases of LLMs and their intersection. We discover significant variance in alignment across languages, challenging the assumption of uniform moral reasoning in AI systems and highlighting the importance of incorporating diverse perspectives in AI ethics. The results underscore the need for further research on the integration of multilingual dimensions in responsible AI research to ensure fair and equitable AI interactions worldwide.

3736. Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs

链接：https://iclr.cc/virtual/2025/poster/31078 abstract： Large Vision-Language Models (LVLMs) often produce responses that misalign with factual information, a phenomenon known as hallucinations. While hallucinations are well-studied, the exact causes behind them remain underexplored. In this paper, we first investigate the root causes of hallucinations in LVLMs. Our findings reveal that existing mitigation techniques primarily reduce hallucinations for visual recognition prompts—those that require simple descriptions of visual elements—but fail for cognitive prompts that demand deliberate reasoning. We identify the core issue as a lack of true visual perception in LVLMs: although they can accurately recognize visual elements, they struggle to fully interpret these elements in the context of the input prompt and effectively link this recognition to their internal knowledge, which is critical for reasoning. To address this gap, we introduce Visual Description Grounded Decoding (VDGD), a simple, robust, and training-free method designed to enhance visual perception and improve reasoning capabilities in LVLMs. VDGD works by first generating a detailed description of the image and appending it as a prefix to the instruction. During response generation, tokens are sampled based on their KL divergence to the description, favoring candidates with lower divergence. Experimental results on multiple visual reasoning benchmarks and LVLMs demonstrate that VDGD consistently outperforms existing baselines 2% - 33%. Finally, we introduce VaLLu, a benchmark designed for comprehensive evaluation of the cognitive capabilities of LVLMs.

3737. Structural-Entropy-Based Sample Selection for Efficient and Effective Learning

链接：https://iclr.cc/virtual/2025/poster/27779 abstract： Sample selection improves the efficiency and effectiveness of machine learning models by providing informative and representative samples. Typically, samples can be modeled as a sample graph, where nodes are samples and edges represent their similarities. Most existing methods are based on local information, such as the training difficulty of samples, thereby overlooking global information, such as connectivity patterns. This oversight can result in suboptimal selection because global information is crucial for ensuring that the selected samples well represent the structural properties of the graph. To address this issue, we employ structural entropy to quantify global information and losslessly decompose it from the whole graph to individual nodes using the Shapley value. Based on the decomposition, we present $\textbf{S}$tructural-$\textbf{E}$ntropy-based sample $\textbf{S}$election ($\textbf{SES}$), a method that integrates both global and local information to select informative and representative samples. SES begins by constructing a kNN-graph among samples based on their similarities. It then measures sample importance by combining structural entropy (global metric) with training difficulty (local metric). Finally, SES applies importance-biased blue noise sampling to select a set of diverse and representative samples. Comprehensive experiments on three learning scenarios --- supervised learning, active learning, and continual learning --- clearly demonstrate the effectiveness of our method.

3738. Diffusion^2: Dynamic 3D Content Generation via Score Composition of Video and Multi-view Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28867 abstract： Recent advancements in 3D generation are predominantly propelled by improvements in 3D-aware image diffusion models. These models are pretrained on Internet-scale image data and fine-tuned on massive 3D data, offering the capability of producing highly consistent multi-view images. However, due to the scarcity of synchronized multi-view video data, it remains challenging to adapt this paradigm to 4D generation directly. Despite that, the available video and 3D data are adequate for training video and multi-view diffusion models separately that can provide satisfactory dynamic and geometric priors respectively. To take advantage of both, this paper presents Diffusion^2, a novel framework for dynamic 3D content creation that reconciles the knowledge about geometric consistency and temporal smoothness from these models to directly sample dense multi-view multi-frame images which can be employed to optimize continuous 4D representation. Specifically, we design a simple yet effective denoising strategy via score composition of pretrained video and multi-view diffusion models based on the probability structure of the target image array. To alleviate the potential conflicts between two heterogeneous scores, we further introduce variance-reducing sampling via interpolated steps, facilitating smooth and stable generation. Owing to the high parallelism of the proposed image generation process and the efficiency of the modern 4D reconstruction pipeline, our framework can generate 4D content within few minutes. Notably, our method circumvents the reliance on expensive and hard-to-scale 4D data, thereby having the potential to benefit from the scaling

of the foundation video and multi-view diffusion models. Extensive experiments demonstrate the efficacy of our proposed framework in generating highly seamless and consistent 4D assets under various types of conditions.

3739. Discriminator-Guided Embodied Planning for LLM Agent

链接：https://iclr.cc/virtual/2025/poster/29523 abstract： Large Language Models (LLMs) have showcased remarkable reasoning capabilities in various domains, yet face challenges in complex embodied tasks due to the need for a coherent long-term policy and context-sensitive environmental understanding. Previous work performed LLM refinement relying on outcome-supervised feedback, which can be costly and ineffective. In this work, we introduce a novel framework, Discriminator-Guided Action Optimization (DGAP), for facilitating the optimization of LLM action plans via step-wise signals. Specifically, we employ a limited set of demonstrations to enable the discriminator to learn a score function, which assesses the alignment between LLM-generated actions and the underlying optimal ones at every step. Based on the discriminator, LLMs are prompted to generate actions that maximize the score, utilizing historical action-score pair trajectories as guidance. Under mild conditions, DGAP resembles critic-regularized optimization and has been demonstrated to achieve a stronger policy than the LLM planner. In experiments across different LLMs (GPT-4, Llama3-70B) in ScienceWorld and VirtualHome, our method achieves superior performance and better efficiency than previous methods.

3740. Controllable Unlearning for Image-to-Image Generative Models via ϵ-Constrained Optimization

链接：https://iclr.cc/virtual/2025/poster/30696 abstract： While generative models have made significant advancements in recent years, they also raise concerns such as privacy breaches and biases. Machine unlearning has emerged as a viable solution, aiming to remove specific training data, e.g., containing private information and bias, from models. In this paper, we study the machine unlearning problem in Image-to-Image (I2I) generative models. Previous studies mainly treat it as a single objective optimization problem, offering a solitary solution, thereby neglecting the varied user expectations towards the trade-off between complete unlearning and model utility. To address this issue, we propose a controllable unlearning framework that uses a control coefficient ϵ to control the trade-off. We reformulate the I2I generative model unlearning problem into a ϵ-constrained optimization problem and solve it with a gradient-based method to find optimal solutions for unlearning boundaries. These boundaries define the valid range for the control coefficient. Within this range, every yielded solution is theoretically guaranteed with Pareto optimality. We also analyze the convergence rate of our framework under various control functions. Extensive experiments on two benchmark datasets across three mainstream I2I models demonstrate the effectiveness of our controllable unlearning framework.

3741. Cyclic Contrastive Knowledge Transfer for Open-Vocabulary Object Detection

链接：https://iclr.cc/virtual/2025/poster/30109 abstract： In pursuit of detecting unstinted objects that extend beyond predefined categories, prior arts of open-vocabulary object detection (OVD) typically resort to pretrained vision-language models (VLMs) for base-to-novel category generalization. However, to mitigate the misalignment between upstream image-text pretraining and downstream region-level perception, additional supervisions are indispensable, e.g., image-text pairs or pseudo annotations generated via self-training strategies. In this work, we propose CCKT-Det trained without any extra supervision. The proposed framework constructs a cyclic and dynamic knowledge transfer from language queries and visual region features extracted from VLMs, which forces the detector to closely align with the visual-semantic space of VLMs. Specifically, 1) we prefilter and inject semantic priors to guide the learning of queries, and 2) introduce a regional contrastive loss to improve the awareness of queries on novel objects. CCKT-Det can consistently improve performance as the scale of VLMs increases, all while requiring the detector at a moderate level of computation overhead. Comprehensive experimental results demonstrate that our method achieves performance gain of +2.9% and +10.2% AP_{50} over previous state-of-the-arts on the challenging COCO benchmark, both without and with a stronger teacher model.

3742. Multi-Resolution Decomposable Diffusion Model for Non-Stationary Time Series Anomaly Detection

链接：https://iclr.cc/virtual/2025/poster/28922 abstract： Recently, generative models have shown considerable promise in unsupervised time series anomaly detection. Nonetheless, the task of effectively capturing complex temporal patterns and minimizing false alarms becomes increasingly challenging when dealing with non-stationary time series, characterized by continuously fluctuating statistical attributes and joint distributions. To confront these challenges, we underscore the benefits of multi-resolution modeling, which improves the ability to distinguish between anomalies and non-stationary behaviors by leveraging correlations across various resolution scales. In response, we introduce a Multi-Resolution Decomposable Diffusion Model (MODEM), which integrates a coarse-to-fine diffusion paradigm with a frequency-enhanced decomposable network to adeptly navigate the intricacies of non-stationarity. Technically, the coarse-to-fine diffusion model embeds cross-resolution correlations into the forward process to optimize diffusion transitions mathematically. It then innovatively employs low-resolution recovery to guide the reverse trajectories of high-resolution series in a coarse-to-fine manner, enhancing the model's ability to learn and elucidate underlying temporal patterns. Furthermore, the frequency-enhanced decomposable network operates in the frequency domain to extract globally shared time-invariant information and time-variant temporal dynamics for accurate series

reconstruction. Extensive experiments conducted across five real-world datasets demonstrate that our proposed MODEM achieves state-of-the-art performance and can be generalized to other time series tasks.

3743. DisPose: Disentangling Pose Guidance for Controllable Human Image Animation

链接：https://iclr.cc/virtual/2025/poster/30606 abstract： Controllable human image animation aims to generate videos from reference images using driving videos. Due to the limited control signals provided by sparse guidance (e.g., skeleton pose), recent works have attempted to introduce additional dense conditions (e.g., depth map) to ensure motion alignment. However, such strict dense guidance impairs the quality of the generated video when the body shape of the reference character differs significantly from that of the driving video. In this paper, we present DisPose to mine more generalizable and effective control signals without additional dense input, which disentangles the sparse skeleton pose in human image animation into motion field guidance and keypoint correspondence. Specifically, we generate a dense motion field from a sparse motion field and the reference image, which provides region-level dense guidance while maintaining the generalization of the sparse pose control. We also extract diffusion features corresponding to pose keypoints from the reference image, and then these point features are transferred to the target pose to provide distinct identity information. To seamlessly integrate into existing models, we propose a plug-and-play hybrid ControlNet that improves the quality and consistency of generated videos while freezing the existing model parameters. Extensive qualitative and quantitative experiments demonstrate the superiority of DisPose compared to current methods. Project page: https://github.com/lihxxx/DisPose.

3744. LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioning

链接：https://iclr.cc/virtual/2025/poster/29989 abstract： Language plays a vital role in the realm of human motion. Existing methods have largely depended on CLIP text embeddings for motion generation, yet they fall short in effectively aligning language and motion due to CLIP’s pretraining on static image-text pairs. This work introduces LaMP, a novel Language-Motion Pretraining model, which transitions from a language-vision to a more suitable language-motion latent space. It addresses key limitations by generating motion-informative text embeddings, significantly enhancing the relevance and semantics of generated motion sequences. With LaMP, we advance three key tasks: text-to-motion generation, motion-text retrieval, and motion captioning through aligned language-motion representation learning. For generation, LaMP instead of CLIP provides the text condition, and an autoregressive masked prediction is designed to achieve mask modeling without rank collapse in transformers. For retrieval, motion features from LaMP’s motion transformer interact with query tokens to retrieve text features from the text transformer, and vice versa. For captioning, we finetune a large language model with the language-informative motion features to develop a strong motion captioning model. In addition, we introduce the LaMP-BertScore metric to assess the alignment of generated motions with textual descriptions. Extensive experimental results on multiple datasets demonstrate substantial improvements over previous methods across all three tasks. Project page: https://aigc3d.github.io/LaMP

3745. VoxDialogue: Can Spoken Dialogue Systems Understand Information Beyond Words?

链接：https://iclr.cc/virtual/2025/poster/27903 abstract： With the rapid advancement of large models, voice assistants are gradually acquiring the ability to engage in open-ended daily conversations with humans. However, current spoken dialogue systems often overlook multi-modal information in audio beyond text, such as speech rate, volume, emphasis, and background sounds. Relying solely on Automatic Speech Recognition (ASR) can lead to the loss of valuable auditory cues, thereby weakening the system’s ability to generate contextually appropriate responses. To address this limitation, we propose \textbf{VoxDialogue}, a comprehensive benchmark for evaluating the ability of spoken dialogue systems to understand multi-modal information beyond text. Specifically, we have identified 12 attributes highly correlated with acoustic information beyond words and have meticulously designed corresponding spoken dialogue test sets for each attribute, encompassing a total of 4.5K multi-turn spoken dialogue samples. Finally, we evaluated several existing spoken dialogue models, analyzing their performance on the 12 attribute subsets of VoxDialogue. Experiments have shown that in spoken dialogue scenarios, many acoustic cues cannot be conveyed through textual information and must be directly interpreted from the audio input. In contrast, while direct spoken dialogue systems excel at processing acoustic signals, they still face limitations in handling complex dialogue tasks due to their restricted context understanding capabilities. All data and code will be open source at \url{https://voxdialogue.github.io/}.

3746. Mixture-of-Agents Enhances Large Language Model Capabilities

链接：https://iclr.cc/virtual/2025/poster/28787 abstract： Recent advances in large language models (LLMs) demonstrate substantial capabilities in natural language understanding and generation tasks. With the growing number of LLMs, how to harness the collective expertise of multiple LLMs is an exciting open direction. Toward this goal, we propose a new approach that leverages the collective strengths of multiple LLMs through a Mixture-of-Agents (MoA) methodology. In our approach, we construct a layered MoA architecture wherein each layer comprises multiple LLM agents. Each agent takes all the outputs from agents in the previous layer as auxiliary information in generating its response. MoA models achieves state-of-art performance on AlpacaEval 2.0, Arena-Hard, MT-Bench, and FLASK, surpassing GPT-4 Omni. For example, our MoA using only open-

source LLMs achieves a score of 65.1% on AlpacaEval 2.0 compared to 57.5% by GPT-4 Omni.

3747. Scaling Instruction-tuned LLMs to Million-token Contexts via Hierarchical Synthetic Data Generation

链接：https://iclr.cc/virtual/2025/poster/30552 abstract： Large Language Models (LLMs) struggle with long-context reasoning, not only due to the quadratic scaling of computational complexity with sequence length but also because of the scarcity and expense of annotating long-context data. There has been barely any open-source work that systematically ablates long-context data, nor is there any openly available instruction tuning dataset with contexts surpassing 100K tokens. To bridge this gap, we introduce a novel post-training synthetic data generation strategy designed to efficiently extend the context window of LLMs while preserving their general task performance. Our approach scalably extends to arbitrarily long context lengths, unconstrained by the length of available real-world data, which effectively addresses the scarcity of raw long-context data. Through a step-by-step rotary position embedding (RoPE) scaling training strategy, we demonstrate that our model, with a context length of up to 1M tokens, performs well on the RULER benchmark and InfiniteBench and maintains robust performance on general language tasks.

3748. Train Small, Infer Large: Memory-Efficient LoRA Training for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28143 abstract： Large Language Models (LLMs) have significantly advanced natural language processing with exceptional task generalization capabilities. Low-Rank Adaption (LoRA) offers a cost-effective fine-tuning solution, freezing the original model parameters and training only lightweight, low-rank adapter matrices. However, the memory footprint of LoRA is largely dominated by the original model parameters. To mitigate this, we propose LoRAM, a memory-efficient LoRA training scheme founded on the intuition that many neurons in over-parameterized LLMs have low training utility but are essential for inference. LoRAM presents a unique twist: it trains on a pruned (small) model to obtain pruned low-rank matrices, which are then recovered and utilized with the original (large) model for inference. Additionally, minimal-cost continual pre-training, performed by the model publishers in advance, aligns the knowledge discrepancy between pruned and original models. Our extensive experiments demonstrate the efficacy of LoRAM across various pruning strategies and downstream tasks. For a model with 70 billion parameters, LoRAM enables training on a GPU with only 20G HBM, replacing an A100-80G GPU for LoRA training and 15 GPUs for full fine-tuning. Specifically, QLoRAM implemented by structured pruning combined with 4-bit quantization, for LLaMA-3.1-70B (LLaMA-2-70B), reduces the parameter storage cost that dominates the memory usage in low-rank matrix training by 15.81× (16.95×), while achieving dominant performance gains over both the original LLaMA-3.1-70B (LLaMA-2-70B) and LoRA-trained LLaMA-3.1-8B (LLaMA-2-13B). Code is available at https://github.com/junzhang-zj/LoRAM.

3749. CLIPDrag: Combining Text-based and Drag-based Instructions for Image Editing

链接：https://iclr.cc/virtual/2025/poster/31150 abstract： Precise and flexible image editing remains a fundamental challenge in computer vision. Based on the modified areas, most editing methods can be divided into two main types: global editing and local editing. In this paper, we choose the two most common editing approaches (\ie text-based editing and drag-based editing) and analyze their drawbacks. Specifically, text-based methods often fail to describe the desired modifications precisely, while drag-based methods suffer from ambiguity. To address these issues, we proposed \textbf{CLIPDrag}, a novel image editing method that is the first to combine text and drag signals for precise and ambiguity-free manipulations on diffusion models. To fully leverage these two signals, we treat text signals as global guidance and drag points as local information. Then we introduce a novel global-local motion supervision method to integrate text signals into existing drag-based methods by adapting a pre-trained language-vision model like CLIP. Furthermore, we also address the problem of slow convergence in CLIPDrag by presenting a fast point-tracking method that enforces drag points moving toward correct directions. Extensive experiments demonstrate that CLIPDrag outperforms existing single drag-based methods or text-based methods.

3750. CollabEdit: Towards Non-destructive Collaborative Knowledge Editing

链接：https://iclr.cc/virtual/2025/poster/31140 abstract：

3751. Bridging Context Gaps: Leveraging Coreference Resolution for Long Contextual Understanding

链接：https://iclr.cc/virtual/2025/poster/29052 abstract： Large language models (LLMs) have shown remarkable capabilities in natural language processing; however, they still face difficulties when tasked with understanding lengthy contexts and executing effective question answering. These challenges often arise due to the complexity and ambiguity present in longer texts. To enhance the performance of LLMs in such scenarios, we introduce the Long Question Coreference Adaptation (LQCA) method. This innovative framework focuses on coreference resolution tailored to long contexts, allowing the model to identify and manage references effectively. The LQCA method encompasses four key steps: resolving coreferences within sub-documents,

computing the distances between mentions, defining a representative mention for coreference, and answering questions through mention replacement. By processing information systematically, the framework provides easier-to-handle partitions for LLMs, promoting better understanding. Experimental evaluations on a range of LLMs and datasets have yielded positive results, with a notable improvements on OpenAI-o1-mini and GPT-4o models, highlighting the effectiveness of leveraging coreference resolution to bridge context gaps in question answering. Our code is public at https://github.com/OceannTwT/LQCA.

3752. Tool-Planner: Task Planning with Clusters across Multiple Tools

链接：https://iclr.cc/virtual/2025/poster/28985 abstract： Large language models (LLMs) have demonstrated exceptional reasoning capabilities, enabling them to solve various complex problems. Recently, this ability has been applied to the paradigm of tool learning. Tool learning involves providing examples of tool usage and their corresponding functions, allowing LLMs to formulate plans and demonstrate the process of invoking and executing each tool. LLMs can address tasks that they cannot complete independently, thereby enhancing their potential across different tasks. However, this approach faces two key challenges. First, redundant error correction leads to unstable planning and long execution time. Additionally, designing a correct plan among multiple tools is also a challenge in tool learning. To address these issues, we propose Tool-Planner, a task-processing framework based on toolkits. Tool-Planner groups tools based on the API functions with the same function into a toolkit and allows LLMs to implement planning across the various toolkits. When a tool error occurs, the language model can reselect and adjust tools based on the toolkit. Experiments show that our approach demonstrates a high pass and win rate across different datasets and optimizes the planning scheme for tool learning in models such as GPT-4 and Claude 3, showcasing the potential of our method. Our code is public at https://github.com/OceannTwT/Tool-Planner.

3753. Small Models are LLM Knowledge Triggers for Medical Tabular Prediction

链接：https://iclr.cc/virtual/2025/poster/29343 abstract： Recent development in large language models (LLMs) has demonstrated impressive domain proficiency on unstructured textual or multi-modal tasks. However, despite with intrinsic world knowledge, their application on structured tabular data prediction still lags behind, primarily due to the numerical insensitivity and modality discrepancy that brings a gap between LLM reasoning and statistical tabular learning. Unlike textual or vision data (e.g., electronic clinical notes or medical imaging data), tabular data is often presented in heterogeneous numerical values (e.g., CBC reports). This ubiquitous data format requires intensive expert annotation, and its numerical nature limits LLMs' capability to effectively transfer untapped domain expertise. In this paper, we propose SERSAL, a general self-prompting method by synergy learning with small models to enhance LLM tabular prediction in an unsupervised manner. Specifically, SERSAL utilizes the LLM's prior outcomes as original soft noisy annotations, which are dynamically leveraged to teach a better small student model. Reversely, the outcomes from the trained small model are used to teach the LLM to further refine its real capability. This process can be repeatedly applied to gradually distill refined knowledge for continuous progress. Comprehensive experiments on widely used medical domain tabular datasets show that, without access to gold labels, applying SERSAL to OpenAI GPT reasoning process attains substantial improvement compared to linguistic prompting methods, which serves as an orthogonal direction for tabular LLM, and increasing prompting bonus is observed as more powerful LLMs appear. Codes are available at https://github.com/jyansir/sersal.

3754. Greener GRASS: Enhancing GNNs with Encoding, Rewiring, and Attention

链接：https://iclr.cc/virtual/2025/poster/28209 abstract： Graph Neural Networks (GNNs) have become important tools for machine learning on graph-structured data. In this paper, we explore the synergistic combination of graph encoding, graph rewiring, and graph attention, by introducing Graph Attention with Stochastic Structures (GRASS), a novel GNN architecture. GRASS utilizes relative random walk probabilities (RRWP) encoding and a novel decomposed variant (D-RRWP) to efficiently capture structural information. It rewires the input graph by superimposing a random regular graph to enhance long-range information propagation. It also employs a novel additive attention mechanism tailored for graph-structured data. Our empirical evaluations demonstrate that GRASS achieves state-of-the-art performance on multiple benchmark datasets, including a 20.3% reduction in mean absolute error on the ZINC dataset.

3755. Graph Transformers Dream of Electric Flow

链接：https://iclr.cc/virtual/2025/poster/28191 abstract： We show theoretically and empirically that the linear Transformer, when applied to graph data, can implement algorithms that solve canonical problems such as electric flow and eigenvector decomposition. The Transformer has access to information on the input graph only via the graph's incidence matrix. We present explicit weight configurations for implementing each algorithm, and we bound the constructed Transformers' errors by the errors of the underlying algorithms. Our theoretical findings are corroborated by experiments on synthetic data. Additionally, on a real-world molecular regression task, we observe that the linear Transformer is capable of learning a more effective positional encoding than the default one based on Laplacian eigenvectors. Our work is an initial step towards elucidating the inner-workings of the Transformer for graph data. Code is available at https://github.com/chengxiang/LinearGraphTransformer

3756. On Statistical Rates of Conditional Diffusion Transformers:

Approximation, Estimation and Minimax Optimality

链接：https://iclr.cc/virtual/2025/poster/29072 abstract： We investigate the approximation and estimation rates of conditional diffusion transformers (DiTs) with classifier-free guidance. We present a comprehensive analysis for “in-context” conditional DiTs under various common assumptions: generic and strong Hölder, linear latent (subspace), and Lipschitz score function assumptions. Importantly, we establish minimax optimality of DiTs by leveraging score function regularity. Specifically, we discretize the input domains into infinitesimal grids and then perform term-by-term Taylor expansions on the conditional diffusion score function under the Hölder smooth data assumption. This enables fine-grained use of transformers’ universal approximation through a more detailed piecewise constant approximation, and hence obtains tighter bounds. Additionally, we extend our analysis to latent settings. Our findings establish statistical limits for DiTs and offer practical guidance toward more efficient and accurate designs.

3757. Learning Equivariant Non-Local Electron Density Functionals

链接：https://iclr.cc/virtual/2025/poster/30333 abstract： The accuracy of density functional theory hinges on the approximation of non-local contributions to the exchange-correlation (XC) functional. To date, machine-learned and human-designed approximations suffer from insufficient accuracy, limited scalability, or dependence on costly reference data. To address these issues, we introduce Equivariant Graph Exchange Correlation (EG-XC), a novel non-local XC functional based on equivariant graph neural networks (GNNs). Where previous works relied on semi-local functionals or fixed-size descriptors of the density, we compress the electron density into an SO(3)-equivariant nuclei-centered point cloud for efficient non-local atomic-range interactions. By applying an equivariant GNN on this point cloud, we capture molecular-range interactions in a scalable and accurate manner. To train EG-XC, we differentiate through a self-consistent field solver requiring only energy targets. In our empirical evaluation, we find EG-XC to accurately reconstruct `gold-standard' CCSD(T) energies on MD17. On out-of-distribution conformations of 3BPA, EG-XC reduces the relative MAE by 35% to 50%. Remarkably, EG-XC excels in data efficiency and molecular size extrapolation on QM9, matching force fields trained on 5 times more and larger molecules. On identical training sets, EG-XC yields on average 51% lower MAEs.

3758. Can Transformers Do Enumerative Geometry?

链接：https://iclr.cc/virtual/2025/poster/31007 abstract： We introduce a Transformer-based approach to computational enumerative geometry, specifically targeting the computation of ψ-class intersection numbers on the moduli space of curves. Traditional methods for calculating these numbers suffer from factorial computational complexity, making them impractical to use. By reformulating the problem as a continuous optimization task, we compute intersection numbers across a wide value range from 10^{-45} to 10^{45}. To capture the recursive nature inherent in these intersection numbers, we propose the Dynamic Range Activator (DRA), a new activation function that enhances the Transformer's ability to model recursive patterns and handle severe heteroscedasticity. Given precision requirements for computing the intersections, we quantify the uncertainty of the predictions using Conformal Prediction with a dynamic sliding window adaptive to the partitions of equivalent number of marked points. To the best of our knowledge, there has been no prior work on modeling recursive functions with such a high-variance and factorial growth. Beyond simply computing intersection numbers, we explore the enumerative "world-model" of Transformers. Our interpretability analysis reveals that the network is implicitly modeling the Virasoro constraints in a purely data-driven manner. Moreover, through abductive hypothesis testing, probing, and causal inference, we uncover evidence of an emergent internal representation of the the large-genus asymptotic of ψ-class intersection numbers. These findings suggest that the network internalizes the parameters of the asymptotic closed-form and the polynomiality phenomenon of ψ-class intersection numbers in a non-linear manner.

3759. No Equations Needed: Learning System Dynamics Without Relying on Closed-Form ODEs

链接：https://iclr.cc/virtual/2025/poster/28568 abstract： Data-driven modeling of dynamical systems is a crucial area of machine learning. In many scenarios, a thorough understanding of the model’s behavior becomes essential for practical applications. For instance, understanding the behavior of a pharmacokinetic model, constructed as part of drug development, may allow us to both verify its biological plausibility (e.g., the drug concentration curve is non-negative and decays to zero in the long term) and to design dosing guidelines (e.g., by looking at the peak concentration and its timing). Discovery of closed-form ordinary differential equations (ODEs) can be employed to obtain such insights by finding a compact mathematical equation and then analyzing it (a two-step approach). However, its widespread use is currently hindered because the analysis process may be time-consuming, requiring substantial mathematical expertise, or even impossible if the equation is too complex. Moreover, if the found equation's behavior does not satisfy the requirements, editing it or influencing the discovery algorithms to rectify it is challenging as the link between the symbolic form of an ODE and its behavior can be elusive. This paper proposes a conceptual shift to modeling low-dimensional dynamical systems by departing from the traditional two-step modeling process. Instead of first discovering a closed-form equation and then analyzing it, our approach, direct semantic modeling, predicts the semantic representation of the dynamical system (i.e., description of its behavior) directly from data, bypassing the need for complex post-hoc analysis. This direct approach also allows the incorporation of intuitive inductive biases into the optimization algorithm and editing the model's behavior directly, ensuring that the model meets the desired specifications. Our approach not only simplifies the modeling pipeline but also enhances the transparency and flexibility of the resulting models compared to traditional closed-form ODEs.

3760. Zero-shot forecasting of chaotic systems

链接：https://iclr.cc/virtual/2025/poster/29513 abstract：

3761. Associative memory and dead neurons

链接：https://iclr.cc/virtual/2025/poster/28453 abstract：

3762. Foundation Models Secretly Understand Neural Network Weights: Enhancing Hypernetwork Architectures with Foundation Models

链接：https://iclr.cc/virtual/2025/poster/29067 abstract： Large pre-trained models, or foundation models, have shown impressive performance when adapted to a variety of downstream tasks, often out-performing specialized models. Hypernetworks, neural networks that generate some or all of the parameters of another neural network, have become an increasingly important technique for conditioning and generalizing implicit neural representations (INRs), which represent signals or objects such as audio or 3D shapes using a neural network. However, despite the potential benefits of incorporating foundation models in hypernetwork methods, this research direction has not been investigated, likely due to the dissimilarity of the weight generation task with other visual tasks. To address this gap, we (1) show how foundation models can improve hypernetworks with Transformer-based architectures, (2) provide an empirical analysis of the benefits of foundation models for hypernetworks through the lens of the generalizable INR task, showing that leveraging foundation models improves performance, generalizability, and data efficiency across a variety of algorithms and modalities. We also provide further analysis in examining the design space of foundation model-based hypernetworks, including examining the choice of foundation models, algorithms, and the effect of scaling foundation models.

3763. Generation and Comprehension Hand-in-Hand: Vision-guided Expression Diffusion for Boosting Referring Expression Generation and Comprehension

链接：https://iclr.cc/virtual/2025/poster/31181 abstract： Referring expression generation (REG) and comprehension (REC) are vital and complementary in joint visual and textual reasoning. Existing REC datasets typically contain insufficient image-expression pairs for training, hindering the generalization of REC models to unseen referring expressions. Moreover, REG methods frequently struggle to bridge the visual and textual domains due to the limited capacity, leading to low-quality and restricted diversity in expression generation. To address these issues, we propose a novel Vlsion-guided Expression Diffusion Model (VIE-DM) for the REG task, where diverse synonymous expressions adhering to both image and text contexts of the target object are generated to augment REC datasets. VIE-DM consists of a vision-text condition (VTC) module and a transformer decoder. Our VTC and token selection design effectively addresses the feature discrepancy problem prevalent in existing REG methods. This enables us to generate high-quality, diverse synonymous expressions that can serve as augmented data for REC model learning. Extensive experiments on five datasets demonstrate the high quality and large diversity of our generated expressions. Furthermore, the augmented image-expression pairs consistently enhance the performance of existing REC models, achieving state-of-the-art results.

3764. 3DGS-Drag: Dragging Gaussians for Intuitive Point-Based 3D Editing

链接：https://iclr.cc/virtual/2025/poster/30830 abstract： The transformative potential of 3D content creation has been progressively unlocked through advancements in generative models. Recently, intuitive drag editing with geometric changes has attracted significant attention in 2D editing yet remains challenging for 3D scenes. In this paper, we introduce 3DGS-Drag, a point-based 3D editing framework that provides efficient, intuitive drag manipulation of real 3D scenes. Our approach bridges the gap between deformation-based and 2D-editing-based 3D editing methods, addressing their limitations to geometry-related content editing. We leverage two key innovations: deformation guidance utilizing 3D Gaussian Splatting for consistent geometric modifications and diffusion guidance for content correction and visual quality enhancement. A progressive editing strategy further supports aggressive 3D drag edits. Our method enables a wide range of edits, including motion change, shape adjustment, inpainting, and content extension. Experimental results demonstrate the effectiveness of 3DGS-Drag in various scenes, achieving state-of-the-art performance in geometry-related 3D content editing. Notably, the editing is efficient, taking 10 to 20 minutes on a single RTX 4090 GPU.

3765. Sketch2Diagram: Generating Vector Diagrams from Hand-Drawn Sketches

链接：https://iclr.cc/virtual/2025/poster/30027 abstract： We address the challenge of automatically generating high-quality vector diagrams from hand-drawn sketches. Vector diagrams are essential for communicating complex ideas across various fields, offering flexibility and scalability. While recent research has progressed in generating diagrams from text descriptions, converting hand-drawn sketches into vector diagrams remains largely unexplored due to the lack of suitable datasets. To address this gap, we introduce SketikZ, a dataset comprising 3,231 pairs of hand-drawn sketches and thier corresponding TikZ

codes as well as reference diagrams.Our evaluations reveal the limitations of state-of-the-art vision and language models (VLMs), positioning SketikZ as a key benchmark for future research in sketch-to-diagram conversion.Along with SketikZ, we present ImgTikZ, an image-to-TikZ model that integrates a 6.7B parameter code-specialized open-source large language model (LLM) with a pre-trained vision encoder. Despite its relatively compact size, ImgTikZ performs comparably to GPT-4o.This success is driven by using our two data augmentation techniques and a multi-candidate inference strategy.Our findings open promising directions for future research in sketch-to-diagram conversion and broader image-to-code generation tasks. SketikZ is publicly available.

3766. Latent Radiance Fields with 3D-aware 2D Representations

链接：https://iclr.cc/virtual/2025/poster/27920 abstract： Latent 3D reconstruction has shown great promise in empowering 3D semantic understanding and 3D generation by distilling 2D features into the 3D space. However, existing approaches struggle with the domain gap between 2D feature space and 3D representations, resulting in degraded rendering performance. To address this challenge, we propose a novel framework that integrates 3D awareness into the 2D latent space. The framework consists of three stages: (1) a correspondence-aware autoencoding method that enhances the 3D consistency of 2D latent representations, (2) a latent radiance field (LRF) that lifts these 3D-aware 2D representations into 3D space, and (3) a VAE-Radiance Field (VAE-RF) alignment strategy that improves image decoding from the rendered 2D representations. Extensive experiments demonstrate that our method outperforms the state-of-the-art latent 3D reconstruction approaches in terms of synthesis performance and cross-dataset generalizability across diverse indoor and outdoor scenes. To our knowledge, this is the first work showing the radiance field representations constructed from 2D latent representations can yield photorealistic 3D reconstruction performance.

3767. Learning Color Equivariant Representations

链接：https://iclr.cc/virtual/2025/poster/30166 abstract： In this paper, we introduce group convolutional neural networks (GCNNs) equivariant to color variation. GCNNs have been designed for a variety of geometric transformations from 2D and 3D rotation groups, to semi-groups such as scale. Despite the improved interpretability, accuracy and generalizability of these architectures, GCNNs have seen limited application in the context of perceptual quantities. Notably, the recent CEConv network uses a GCNN to achieve equivariance to hue transformations by convolving input images with a hue rotated RGB filter. However, this approach leads to invalid RGB values which break equivariance and degrade performance. We resolve these issues with a lifting layer that transforms the input image directly, thereby circumventing the issue of invalid RGB values and improving equivariance error by over three orders of magnitude. Moreover, we extend the notion of color equivariance to include equivariance to saturation and luminance shift. Our hue-, saturation-, luminance- and color-equivariant networks achieve strong generalization to out-of-distribution perceptual variations and improved sample efficiency over conventional architectures. We demonstrate the utility of our approach on synthetic and real world datasets where we consistently outperform competitive baselines.

3768. CHAMP: Conformalized 3D Human Multi-Hypothesis Pose Estimators

链接：https://iclr.cc/virtual/2025/poster/28587 abstract： We introduce CHAMP, a novel method for learning sequence-to-sequence, multi-hypothesis 3D human poses from 2D keypoints by leveraging a conditional distribution with a diffusion model. To predict a single output 3D pose sequence, we generate and aggregate multiple 3D pose hypotheses. For better aggregation results, we develop a method to score these hypotheses during training, effectively integrating conformal prediction into the learning process. This process results in a differentiable conformal predictor that is trained end-to-end with the 3D pose estimator. Post-training, the learned scoring model is used as the conformity score, and the 3D pose estimator is combined with a conformal predictor to select the most accurate hypotheses for downstream aggregation. Our results indicate that using a simple mean aggregation on the conformal prediction-filtered hypotheses set yields competitive results. When integrated with more sophisticated aggregation techniques, our method achieves state-of-the-art performance across various metrics and datasets while inheriting the probabilistic guarantees of conformal prediction.

3769. RESfM: Robust Deep Equivariant Structure from Motion

链接：https://iclr.cc/virtual/2025/poster/27825 abstract： Multiview Structure from Motion is a fundamental and challenging computer vision problem. A recent deep-based approach utilized matrix equivariant architectures for simultaneous recovery of camera pose and 3D scene structure from large image collections. That work, however, made the unrealistic assumption that the point tracks given as input are almost clean of outliers. Here, we propose an architecture suited to dealing with outliers by adding a multiview inlier/outlier classification module that respects the model equivariance and by utilizing a robust bundle adjustment step. Experiments demonstrate that our method can be applied successfully in realistic settings that include large image collections and point tracks extracted with common heuristics that include many outliers, achieving state-of-the-art accuracies in almost all runs, superior to existing deep-based methods and on-par with leading classical (non-deep) sequential and global methods.

3770. Empowering Users in Digital Privacy Management through Interactive LLM-Based Agents

链接：https://iclr.cc/virtual/2025/poster/30352 abstract： This paper presents a novel application of large language models (LLMs) to enhance user comprehension of privacy policies through an interactive dialogue agent. We demonstrate that LLMs significantly outperform traditional models in tasks like Data Practice Identification, Choice Identification, Policy Summarization, and Privacy Question Answering, setting new benchmarks in privacy policy analysis. Building on these findings, we introduce an innovative LLM-based agent that functions as an expert system for processing website privacy policies, guiding users through complex legal language without requiring them to pose specific questions. A user study with 100 participants showed that users assisted by the agent had higher comprehension levels (mean score of 2.6 out of 3 vs. 1.8 in the control group), reduced cognitive load (task difficulty ratings of 3.2 out of 10 vs. 7.8), increased confidence in managing privacy, and completed tasks in less time (5.5 minutes vs. 15.8 minutes). This work highlights the potential of LLM-based agents to transform user interaction with privacy policies, leading to more informed consent and empowering users in the digital services landscape.

3771. Smoothing the Shift: Towards Stable Test-Time Adaptation under Complex Multimodal Noises

链接：https://iclr.cc/virtual/2025/poster/28200 abstract：

3772. BP-Modified Local Loss for Efficient Training of Deep Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29913 abstract：

3773. Basis Sharing: Cross-Layer Parameter Sharing for Large Language Model Compression

链接：https://iclr.cc/virtual/2025/poster/28800 abstract：

3774. Are Large Vision Language Models Good Game Players?

链接：https://iclr.cc/virtual/2025/poster/29074 abstract： Large Vision Language Models (LVLMs) have demonstrated remarkable abilities in understanding and reasoning about both visual and textual information. However, existing evaluation methods for LVLMs, primarily based on benchmarks like Visual Question Answering and image captioning, often fail to capture the full scope of LVLMs' capabilities. These benchmarks are limited by issues such as inadequate assessment of detailed visual perception, data contamination, and a lack of focus on multi-turn reasoning. To address these challenges, we propose LVLM-Playground, a game-based evaluation framework designed to provide a comprehensive assessment of LVLMs' cognitive and reasoning skills in structured environments. LVLM-Playground uses a set of games to evaluate LVLMs on four core tasks: Perceiving, Question Answering, Rule Following, and End-to-End Playing, with each target task designed to assess specific abilities, including visual perception, reasoning, decision-making, etc. Based on this framework, we conduct extensive experiments that explore the limitations of current LVLMs, such as handling long structured outputs and perceiving detailed and dense elements. Code and data are publicly available at https://github.com/xinke-wang/LVLM-Playground.

3775. Neural Phylogeny: Fine-Tuning Relationship Detection among Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28610 abstract： Given a collection of neural networks, can we determine which are parent models and which are child models fine-tuned from the parents?In this work, we strive to answer this questionvia introducing a new task termed as neural phylogeny detection, aimed at identifying the existence and direction of the fine-tuning relationship. Specifically, neural phylogeny detection attempts to identify all parent-child model pairs and determine, within each pair, which model is the parent and which is the child.We present two approaches for neural phylogeny detection: a learning-free method and a learning-based method. First, we propose a metric that leverages the distance from network parameters to a fake initialization to infer fine-tuning directions. By integrating this metric with traditional clustering algorithms, we propose a series of efficient, learning-free neural phylogeny detection methods. Second, we introduce a transformer-based neural phylogeny detector, which significantly enhances detection accuracy through a learning-based manner. Extensive experiments, ranging from shallow fully-connected networks to open-sourced Stable Diffusion and LLaMA models, progressively validate the effectiveness of both methods. The results demonstrate the reliability of both the learning-free and the learning-based approaches across various learning tasks and network architectures, as well as their ability to detect cross-generational phylogeny between ancestor models and their fine-tuned descendants.

3776. Your Mixture-of-Experts LLM Is Secretly an Embedding Model for Free

链接：https://iclr.cc/virtual/2025/poster/28939 abstract： While large language models (LLMs) excel on generation tasks, their decoder-only architecture often limits their potential as embedding models if no further representation finetuning is applied. Does this contradict their claim of generalists? To answer the question, we take a closer look at Mixture-of-Experts (MoE) LLMs. Our study shows that the expert routers in MoE LLMs can serve as an off-the-shelf embedding model with promising performance on a diverse class of embedding-focused tasks, without requiring any finetuning. Moreover, our extensive analysis

shows that the MoE routing weights (RW) is complementary to the hidden state (HS) of LLMs, a widely-used embedding. Compared to HS, we find that RW is more robust to the choice of prompts and focuses on high-level semantics. Motivated by the analysis, we propose MoEE combining RW and HS, which achieves better performance than using either separately. Our exploration of their combination and prompting strategy shed several novel insights, e.g., a weighted sum of RW and HS similarities outperforms the similarity on their concatenation. Our experiments are conducted on 6 embedding tasks with 20 datasets from the Massive Text Embedding Benchmark (MTEB). The results demonstrate the significant improvement brought by MoEE to LLM-based embedding without further finetuning.

3777. DOCS: Quantifying Weight Similarity for Deeper Insights into Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29315 abstract： We introduce a novel index, the Distribution of Cosine Similarity (DOCS), for quantitatively assessing the similarity between weight matrices in Large Language Models (LLMs), aiming to facilitate the analysis of their complex architectures. Leveraging DOCS, our analysis uncovers intriguing patterns in the latest open-source LLMs: adjacent layers frequently exhibit high weight similarity and tend to form clusters, suggesting depth-wise functional specialization. Additionally, we prove that DOCS is theoretically effective in quantifying similarity for orthogonal matrices, a crucial aspect given the prevalence of orthogonal initializations in LLMs. This research contributes to a deeper understanding of LLM architecture and behavior, offering tools with potential implications for developing more efficient and interpretable models.

3778. BadJudge: Backdoor Vulnerabilities of LLM-As-A-Judge

链接：https://iclr.cc/virtual/2025/poster/28942 abstract： This paper proposes a novel backdoor threat attacking the LLM-as-a-Judge evaluation regime, where the adversary controls both the candidate and evaluator model. The backdoored evaluator victimizes benign users by unfairly assigning inflated scores to adversary. A trivial single token backdoor poisoning 1% of the evaluator training data triples the adversary's score with respect to their legitimate score. We systematically categorize levels of data access corresponding to three real-world settings, (1) web poisoning, (2) malicious annotator, and (3) weight poisoning. These regimes reflect a weak to strong escalation of data access that highly correlates with attack severity. Under the weakest assumptions - web poisoning (1), the adversary still induces a 20% score inflation. Likewise, in the (3) weight poisoning regime, the stronger assumptions enable the adversary to inflate their scores from 1.5/5 to 4.9/5. The backdoor threat generalizes across different evaluator architectures, trigger designs, evaluation tasks, and poisoning rates. By poisoning 10% of the evaluator training data, we control toxicity judges (Guardrails) to misclassify toxic prompts as non-toxic 89% of the time, and document reranker judges in RAG to rank the poisoned document first 97% of the time. LLM-as-a-Judge is uniquely positioned at the intersection of ethics and technology, where social implications of mislead model selection and evaluation constrain the available defensive tools. Amidst these challenges, model merging emerges as a principled tool to offset the backdoor, reducing ASR to near 0% whilst maintaining SOTA performance. Model merging's low computational cost and convenient integration into the current LLM Judge training pipeline position it as a promising avenue for backdoor mitigation in the LLM-as-a-Judge setting.

3779. Improving Neural Network Accuracy by Concurrently Training with a Twin Network

链接：https://iclr.cc/virtual/2025/poster/29553 abstract： Recently within Spiking Neural Networks, a method called Twin Network Augmentation (TNA) has been introduced. This technique claims to improve the validation accuracy of a Spiking Neural Network simply by training two networks in conjunction and matching the logits via the Mean Squared Error loss. In this paper, we validate the viability of this method on a wide range of popular Convolutional Neural Network (CNN) benchmarks and compare this approach to existing Knowledge Distillation schemes. Next, we conduct a in-depth study of the different components that make up TNA and determine that its effectiveness is not solely situated in an increase of trainable parameters, but rather the effect of the training methodology. Finally, we analyse the representations learned by networks trained with TNA and highlight their superiority in a number of tasks, thus proving empirically the applicability of Twin Network Augmentation on CNN models.

3780. Democratic Training Against Universal Adversarial Perturbations

链接：https://iclr.cc/virtual/2025/poster/31018 abstract： Despite their advances and success, real-world deep neural networks are known to be vulnerable to adversarial attacks. Universal adversarial perturbation, an input-agnostic attack, poses a serious threat for them to be deployed in security-sensitive systems. In this case, a single universal adversarial perturbation deceives the model on a range of clean inputs without requiring input-specific optimization, which makes it particularly threatening. In this work, we observe that universal adversarial perturbations usually lead to abnormal entropy spectrum in hidden layers, which suggests that the prediction is dominated by a small number of ``feature'' in such cases (rather than democratically by many features). Inspired by this, we propose an efficient yet effective defense method for mitigating UAPs called \emph{Democratic Training} by performing entropy-based model enhancement to suppress the effect of the universal adversarial perturbations in a given model. \emph{Democratic Training} is evaluated with 7 neural networks trained on 5 benchmark datasets and 5 types of state-of-the-art universal adversarial attack methods. The results show that it effectively reduces the attack success rate, improves model robustness and preserves the model accuracy on clean samples.

3781. Severing Spurious Correlations with Data Pruning

链接：https://iclr.cc/virtual/2025/poster/30555 abstract： Deep neural networks have been shown to learn and rely on spurious correlations present in the data that they are trained on. Reliance on such correlations can cause these networks to malfunction when deployed in the real world, where these correlations may no longer hold. To overcome the learning of and reliance on such correlations, recent studies propose approaches that yield promising results. These works, however, study settings where the strength of the spurious signal is significantly greater than that of the core, invariant signal, making it easier to detect the presence of spurious features in individual training samples and allow for further processing. In this paper, we identify new settings where the strength of the spurious signal is relatively weaker, making it difficult to detect any spurious information while continuing to have catastrophic consequences. We also discover that spurious correlations are learned primarily due to only a handful of all the samples containing the spurious feature and develop a novel data pruning technique that identifies and prunes small subsets of the training data that contain these samples. Our proposed technique does not require inferred domain knowledge, information regarding the sample-wise presence or nature of spurious information, or human intervention. Finally, we show that such data pruning attains state-of-the-art performance on previously studied settings where spurious information is identifiable.

3782. Provably Reliable Conformal Prediction Sets in the Presence of Data Poisoning

链接：https://iclr.cc/virtual/2025/poster/28340 abstract： Conformal prediction provides model-agnostic and distribution-free uncertainty quantification through prediction sets that are guaranteed to include the ground truth with any user-specified probability. Yet, conformal prediction is not reliable under poisoning attacks where adversaries manipulate both training and calibration data, which can significantly alter prediction sets in practice. As a solution, we propose reliable prediction sets (RPS): the first efficient method for constructing conformal prediction sets with provable reliability guarantees under poisoning. To ensure reliability under training poisoning, we introduce smoothed score functions that reliably aggregate predictions of classifiers trained on distinct partitions of the training data. To ensure reliability under calibration poisoning, we construct multiple prediction sets, each calibrated on distinct subsets of the calibration data. We then aggregate them into a majority prediction set, which includes a class only if it appears in a majority of the individual sets. Both proposed aggregations mitigate the influence of datapoints in the training and calibration data on the final prediction set. We experimentally validate our approach on image classification tasks, achieving strong reliability while maintaining utility and preserving coverage on clean data. Overall, our approach represents an important step towards more trustworthy uncertainty quantification in the presence of data poisoning.

3783. The "Law" of the Unconscious Contrastive Learner: Probabilistic Alignment of Unpaired Modalities

链接：https://iclr.cc/virtual/2025/poster/30439 abstract：

3784. Bayesian Treatment of the Spectrum of the Empirical Kernel in (Sub)Linear-Width Neural Networks

链接：https://iclr.cc/virtual/2025/poster/29844 abstract：

3785. Oscillatory State-Space Models

链接：https://iclr.cc/virtual/2025/poster/30278 abstract：

3786. Two Sparse Matrices are Better than One: Sparsifying Neural Networks with Double Sparse Factorization

链接：https://iclr.cc/virtual/2025/poster/30435 abstract：

3787. Designing Concise ConvNets with Columnar Stages

链接：https://iclr.cc/virtual/2025/poster/27642 abstract：

3788. Sharpness-Aware Minimization: General Analysis and Improved Rates

链接：https://iclr.cc/virtual/2025/poster/30738 abstract： Sharpness-Aware Minimization (SAM) has emerged as a powerful method for improving generalization in machine learning models by minimizing the sharpness of the loss landscape. However, despite its success, several important questions regarding the convergence properties of SAM in non-convex settings are still

open, including the benefits of using normalization in the update rule, the dependence of the analysis on the restrictive bounded variance assumption, and the convergence guarantees under different sampling strategies. To address these questions, in this paper, we provide a unified analysis of SAM and its unnormalized variant (USAM) under one single flexible update rule (Unified SAM), and we present convergence results of the new algorithm under a relaxed and more natural assumption on the stochastic noise. Our analysis provides convergence guarantees for SAM under different step size selections for non-convex problems and functions that satisfy the Polyak-Lojasiewicz (PL) condition (a non-convex generalization of strongly convex functions). The proposed theory holds under the arbitrary sampling paradigm, which includes importance sampling as special case, allowing us to analyze variants of SAM that were never explicitly considered in the literature. Experiments validate the theoretical findings and further demonstrate the practical effectiveness of Unified SAM in training deep neural networks for image classification tasks.

3789. Local convergence of simultaneous min-max algorithms to differential equilibrium on Riemannian manifold

链接：https://iclr.cc/virtual/2025/poster/29652 abstract： We study min-max algorithms to solve zero-sum differential games onRiemannian manifold.Based on the notions ofdifferential Stackelberg equilibriumand differential Nash equilibrium on Riemannian manifold,we analyze the local convergence of two representative deterministic simultaneous algorithms τ-GDA and τ-SGAto such equilibria.Sufficient conditions are obtained to establish the linear convergence rateof τ-GDA based on the Ostrowski theorem on manifold and spectral analysis. To avoid strong rotational dynamics in τ-GDA, τ-SGA is extended fromthe symplectic gradient-adjustment method in Euclidean space.We analyze an asymptotic approximation of τ-SGA when the learning rate ratio τ is big. In some cases, it can achieve a faster convergence rate to differential Stackelberg equilibrium compared to τ-GDA. We show numerically how the insights obtained from theconvergence analysis may improvethe training of orthogonal Wasserstein GANs using stochastic τ-GDA and τ-SGA on simple benchmarks.

3790. Policy Optimization under Imperfect Human Interactions with Agent-Gated Shared Autonomy

链接：https://iclr.cc/virtual/2025/poster/29983 abstract： We introduce AGSA, an Agent-Gated Shared Autonomy framework that learns from high-level human feedback to tackle the challenges of reward-free training, safe exploration, and imperfect low-level human control. Recent human-in-the loop learning methods enable human participants to intervene a learning agent's control and provide online demonstrations. Nonetheless, these methods rely heavily on perfect human interactions, including accurate human-monitored intervention decisions and near-optimal human demonstrations. AGSA employs a dedicated gating agent to determine when to switch control, thereby reducing the need of constant human monitoring. To obtain a precise and foreseeable gating agent, AGSA trains a long-term gating value function from human evaluative feedback on the gating agent's intervention requests and preference feedback on pairs of human intervention trajectories. Instead of relying on potentially suboptimal human demonstrations, the learning agent is trained using control-switching signals from the gating agent. We provide theoretical insights on performance bounds that respectively describe the ability of the two agents. Experiments are conducted with both simulated and real human participants at different skill levels in challenging continuous control environments. Comparative results highlight that AGSA achieves significant improvements over previous human-in-the-loop learning methods in terms of training safety, policy performance, and user-friendliness.

3791. Simple, Good, Fast: Self-Supervised World Models Free of Baggage

链接：https://iclr.cc/virtual/2025/poster/27740 abstract： What are the essential components of world models? How far do we get with world models that are not employing RNNs, transformers, discrete representations, and image reconstructions? This paper introduces SGF, a Simple, Good, and Fast world model that uses self-supervised representation learning, captures short-time dependencies through frame and action stacking, and enhances robustness against model errors through data augmentation. We extensively discuss SGF's connections to established world models, evaluate the building blocks in ablation studies, and demonstrate good performance through quantitative comparisons on the Atari 100k benchmark. The code is available at https://github.com/jrobine/sgf.

3792. CBMA: Improving Conformal Prediction through Bayesian Model Averaging

链接：https://iclr.cc/virtual/2025/poster/30578 abstract： Conformal prediction has emerged as a popular technique for facilitating valid predictive inference across a spectrum of machine learning models, under minimal assumption of exchangeability. Recently, Hoff (2023) showed that full conformal Bayes provides the most efficient prediction sets (smallest by expected volume) among all prediction sets that are valid at the $(1 - \alpha)$ level if the model is correctly specified. However, a critical issue arises when the Bayesian model itself may be mis-specified, resulting in prediction interval that might be suboptimal, even though it still enjoys the frequentist coverage guarantee. To address this limitation, we propose an innovative solution that combines Bayesian model averaging (BMA) with conformal prediction. This hybrid not only leverages the strengths of Bayesian conformal prediction but also introduces a layer of robustness through model averaging. Theoretically, we prove that the resulting prediction interval will converge to the optimal level of efficiency, if the true model is included among the candidate

models. This assurance of optimality, even under potential model uncertainty, provides a significant improvement over existing methods, ensuring more reliable and precise uncertainty quantification.

3793. End-to-end Learning of Gaussian Mixture Priors for Diffusion Sampler

链接：https://iclr.cc/virtual/2025/poster/28690 abstract： Diffusion models optimized via variational inference (VI) have emerged as a promising tool for generating samples from unnormalized target densities. These models create samples by simulating a stochastic differential equation, starting from a simple, tractable prior, typically a Gaussian distribution. However, when the support of this prior differs greatly from that of the target distribution, diffusion models often struggle to explore effectively or suffer from large discretization errors. Moreover, learning the prior distribution can lead to mode-collapse, exacerbated by the mode-seeking nature of reverse Kullback-Leibler divergence commonly used in VI.To address these challenges, we propose end-to-end learnable Gaussian mixture priors (GMPs). GMPs offer improved control over exploration, adaptability to target support, and increased expressiveness to counteract mode collapse. We further leverage the structure of mixture models by proposing a strategy to iteratively refine the model through the addition of mixture components during training. Our experimental results demonstrate significant performance improvements across a diverse range of real-world and synthetic benchmark problems when using GMPs without requiring additional target evaluations.

3794. Connecting Federated ADMM to Bayes

链接：https://iclr.cc/virtual/2025/poster/28671 abstract： We provide new connections between two distinct federated learning approaches based on (i) ADMM and (ii) Variational Bayes (VB), and propose new variants by combining their complementary strengths. Specifically, we show that the dual variables in ADMM naturally emerge through the "site" parameters used in VB with isotropic Gaussian covariances. Using this, we derive two versions of ADMM from VB that use flexible covariances and functional regularisation, respectively. Through numerical experiments, we validate the improvements obtained in performance. The work shows connection between two fields that are believed to be fundamentally different and combines them to improve federated learning.

3795. Training One-Dimensional Graph Neural Networks is NP-Hard

链接：https://iclr.cc/virtual/2025/poster/30841 abstract： We initiate the study of the computational complexity of training graph neural networks (GNNs). We consider the classical node classification setting; there, the intractability of training multidimensonal GNNs immediately follows from known lower bounds for training classical neural networks (and holds even for trivial GNNs). However, one-dimensional GNNs form a crucial case of interest: the computational complexity of training such networks depends on both the graphical structure of the network and the properties of the involved activation and aggregation functions. As our main result, we establish the NP-hardness of training ReLU-activated one-dimensional GNNs via a highly non-trivial reduction. We complement this result with algorithmic upper bounds for the training problem in the ReLU-activated and linearly-activated settings.

3796. On the Optimal Memorization Capacity of Transformers

链接：https://iclr.cc/virtual/2025/poster/29483 abstract： Recent research in the field of machine learning has increasingly focused on the memorization capacity of Transformers, but how efficient they are is not yet well understood.We demonstrate that Transformers can memorize labels with $\tilde{O}(\sqrt{N})$ parameters in a next-token prediction setting for N input sequences of length n, which is proved to be optimal up to logarithmic factors.This indicates that Transformers can efficiently perform memorization with little influence from the input length n owing to the benefit of parameter sharing.We also analyze the memorization capacity in the sequence-to-sequence setting, and find that $\tilde{O}(\sqrt{nN})$ parameters are not only sufficient, but also necessary at least for Transformers with hardmax.These results suggest that while self-attention mechanisms can efficiently identify input sequences, the feed-forward network becomes a bottleneck when associating a label to each token.

3797. Efficient Online Pruning and Abstraction for Imperfect Information Extensive-Form Games

链接：https://iclr.cc/virtual/2025/poster/29930 abstract： Efficiently computing approximate equilibrium strategies in large Imperfect Information Extensive-Form Games (IIEFGs) poses significant challenges due to the game tree's exponential growth. While pruning and abstraction techniques are essential for complexity reduction, existing methods face two key limitations: (i) Seamless integration of pruning with Counterfactual Regret Minimization (CFR) is nontrivial, and (ii) Pruning and abstraction approaches incur prohibitive computational costs, hindering real-world deployment. We propose Expected-Value Pruning and Abstraction (EVPA), a novel online framework that addresses these challenges through three synergistic components: (i) Expected value estimation using approximate Nash equilibrium strategies to quantify information set utilities, (ii) Minimax pruning before CFR to eliminate a large number of sub-optimal actions permanently, and (iii) Dynamic online information abstraction merging information sets based on their current and future expected values in subgames. Experiments on Heads-up No-Limit Texas Hold'em (HUNL) show EVPA outperforms DeepStack's replication and Slumbot with significant win-rate margins in multiple settings. Remarkably, EVPA requires only 1\%-2\% of the solving time to reach an approximate Nash equilibrium compared to DeepStack's replication.

3798. Strategic Classification With Externalities

链接：https://iclr.cc/virtual/2025/poster/28377 abstract： We propose a new variant of the strategic classification problem: a principal reveals a classifier, and n agents report their (possibly manipulated) features to be classified. Motivated by real-world applications, our model crucially allows the manipulation of one agent to affect another; that is, it explicitly captures inter-agent externalities. The principal-agent interactions are formally modeled as a Stackelberg game, with the resulting agent manipulation dynamics captured as a simultaneous game. We show that under certain assumptions, the pure Nash Equilibrium of this agent manipulation game is unique and can be efficiently computed. Leveraging this result, PAC learning guarantees are established for the learner: informally, we show that it is possible to learn classifiers that minimize loss on the distribution, even when a random number of agents are manipulating their way to a pure Nash Equilibrium. We also comment on the optimization of such classifiers through gradient-based approaches. This work sets the theoretical foundations for a more realistic analysis of classifiers that are robust against multiple strategic actors interacting in a common environment.

3799. Bounds on L_p Errors in Density Ratio Estimation via f-Divergence Loss Functions

链接：https://iclr.cc/virtual/2025/poster/28015 abstract： Density ratio estimation (DRE) is a core technique in machine learning used to capture relationships between two probability distributions. f-divergence loss functions, which are derived from variational representations of f-divergence, have become a standard choice in DRE for achieving cutting-edge performance. This study provides novel theoretical insights into DRE by deriving upper and lower bounds on the L_p errors through f-divergence loss functions. These bounds apply to any estimator belonging to a class of Lipschitz continuous estimators, irrespective of the specific f-divergence loss function employed.The derived bounds are expressed as a product involving the data dimensionality and the expected value of the density ratio raised to the p-th power.Notably, the lower bound includes an exponential term that depends on the Kullback--Leibler (KL) divergence, revealing that the L_p error increases significantly as the KL divergence grows when $p > 1$. This increase becomes even more pronounced as the value of p grows. The theoretical insights are validated through numerical experiments.

3800. Conservative Contextual Bandits: Beyond Linear Representations

链接：https://iclr.cc/virtual/2025/poster/29601 abstract： Conservative Contextual Bandits (CCBs) address safety in sequential decision making by requiring that an agent's policy, along with minimizing regret, also satisfies a safety constraint: the performance is not worse than a baseline policy (e.g., the policy that the company has in production) by more than $(1+\alpha)$ factor. Prior work developed UCB-stylealgorithms for this problem in the multi-armed (Wu et al., 2016) and contextuallinear (Kazerouni et al., 2017) settings.However, in practice the cost of the armsis often a non-linear function, and therefore existing UCB algorithms are ineffective in such settings. In this paper, we consider CCBs beyond the linear case and develop two algorithms $\mathtt{C\text{-}SquareCB}$ and $\mathtt{C\text{-}FastCB}$, using Inverse Gap Weighting (IGW) based exploration and an online regression oracle. We show that the safety constraint is satisfied in high probability and that the regret for $\mathtt{C\text{-}SquareCB}$ is sub-linear in horizon T, while the the regret for $\mathtt{C\text{-}FastCB}$ is first-order and is sub-linear in L^*, the cumulative loss of the optimal policy. Subsequently, we use a neural network for function approximation and online gradient descent as the regression oracle to provide $\tilde{\mathcal{O}}\big(\sqrt{KT} + K\alpha\big)$ and $\tilde{\mathcal{O}}\big(\sqrt{KL^*} + K (1 + 1/\alpha)\big)$ regret bounds respectively. Finally, we demonstrate the efficacy of our algorithms on real world data, and show that they significantly outperform the existing baseline while maintaining the performance guarantee.