401. Federated Granger Causality Learning For Interdependent Clients With State Space Representation

链接：https://iclr.cc/virtual/2025/poster/30056 abstract： Advanced sensors and IoT devices have improved the monitoring and control of complex industrial enterprises. They have also created an interdependent fabric of geographically distributed process operations (clients) across these enterprises. Granger causality is an effective approach to detect and quantify interdependencies by examining how the state of one client affects the states of others over time. Understanding these interdependencies helps capture how localized events, such as faults and disruptions, can propagate throughout the system, potentially leading to widespread operational impacts. However, the large volume and complexity of industrial data present significant challenges in effectively modeling these interdependencies. This paper develops a federated approach to learning Granger causality. We utilize a linear state space system framework that leverages low-dimensional state estimates to analyze interdependencies. This helps address bandwidth limitations and the computational burden commonly associated with centralized data processing. We propose augmenting the client models with the Granger causality information learned by the server through a MachineLearning (ML) function. We examine the co-dependence between the augmented client and server models and reformulate the framework as a standalone ML algorithm providing conditions for its sublinear and linear convergence rates. We also study the convergence of the framework to a centralized oracle model. Moreover, we include a differential privacy analysis to ensure data security while preserving causal insights. Using synthetic data, we conduct comprehensive experiments to demonstrate the robustness of our approach to perturbations in causality, the scalability to the size of communication, number of clients, and the dimensions of raw data. We also evaluate the performance on two real-world industrial control system datasets by reporting the volume of data saved by decentralization.

402. MambaPEFT: Exploring Parameter-Efficient Fine-Tuning for Mamba

链接：https://iclr.cc/virtual/2025/poster/29486 abstract： An ecosystem of Transformer-based models has been established by building large models with extensive data. Parameter-efficient fine-tuning (PEFT) is a crucial technology for deploying these models to downstream tasks with minimal cost while achieving effective performance. Recently, Mamba, a State Space Model (SSM)-based model, has attracted attention as a potential alternative to Transformers. While many large-scale Mamba-based models have been proposed, efficiently adapting pre-trained Mamba-based models to downstream tasks remains unexplored.In this paper, we conduct an exploratory analysis of PEFT methods for Mamba. We investigate the effectiveness of existing PEFT methods for Transformers when applied to Mamba. We also modify these methods to better align with the Mamba architecture. Additionally, we propose new Mamba-specific PEFT methods that leverage the distinctive structure of Mamba. Our experiments indicate that PEFT performs more effectively for Mamba than Transformers. Lastly, we demonstrate how to effectively combine multiple PEFT methods and provide a framework that outperforms previous works. To ensure reproducibility, we will release the code after publication.

403. No Need to Talk: Asynchronous Mixture of Language Models

链接：https://iclr.cc/virtual/2025/poster/28307 abstract： We introduce SMALLTALK LM, an innovative method for training a mixture of language models in an almost asynchronous manner. Eachmodel of the mixture specializes in distinct parts of the data distribution, without the need of high-bandwidth communication between the nodes training each model. At inference, a lightweight router directs a given sequence to a single expert, according to a short prefix. This inference scheme naturally uses a fraction of the parameters from the overall mixture model. Unlike prior works on asynchronous LLM training, our routing method does not rely on full corpus clustering or access to metadata, making it more suitable for real-world applications. Our experiments on language modeling demonstrate that SMALLTALK LM achieves significantly lower perplexity than dense model baselines for the same total training FLOPs and an almost identical inference cost. Finally, in our downstream evaluations we outperform the dense baseline on 75% of the tasks.

404. BingoGuard: LLM Content Moderation Tools with Risk Levels

链接：https://iclr.cc/virtual/2025/poster/30227 abstract： Malicious content generated by large language models (LLMs) can pose varying degrees of harm. Although existing LLM-based moderators can detect harmful content, they struggle to assess risk levels and may miss lower-risk outputs. Accurate risk assessment allows platforms with different safety thresholds to tailor content filtering and rejection. In this paper, we introduce per-topic severity rubrics for 11 harmful topics and build BingoGuard, an LLM-based moderation system designed to predict both binary safety labels and severity levels. To address the lack of annotations on levels of severity, we propose a scalable generate-then-filter framework that first generates responses across different severity levels and then filters out low-quality responses. Using this framework, we create BingoGuardTrain, a training dataset with 54,897 examples covering a variety of topics, response severity, styles, and BingoGuardTest, a test set with 988 examples explicitly labeled based on our severity rubrics that enables fine-grained analysis on model behaviors on different severity levels. Our BingoGuard-8B, trained on BingoGuardTrain, achieves the state-of-the-art performance on several moderation benchmarks, including WildGuardTest and HarmBench, as well as BingoGuardTest, outperforming best public models, WildGuard, by 4.3\%. Our analysis demonstrates that incorporating severity levels into training significantly enhances detection performance and enables the model to effectively gauge the severity of harmful responses. Warning: this paper includes red-teaming examples that may be harmful in nature.

405. HyPoGen: Optimization-Biased Hypernetworks for Generalizable Policy

Generation

链接：https://iclr.cc/virtual/2025/poster/30522 abstract： Policy learning through behavior cloning poses significant challenges, particularly when demonstration data is limited. In this work, we present HyPoGen, a novel optimization-biased hypernetwork for policy generation. The proposed hypernetwork learns to synthesize optimal policy parameters solely from task specifications -- without accessing training data -- by modeling policy generation as an approximation of the optimization process executed over a finite number of steps and assuming these specifications serve as a sufficient representation of the demonstration data. By incorporating structural designs that bias the hypernetwork towards optimization, we can improve its generalization capability while only training on source task demonstrations. During the feed-forward prediction pass, the hypernetwork effectively performs an optimization in the latent (compressed) policy space, which is then decoded into policy parameters for action prediction. Experimental results on locomotion and manipulation benchmarks show that HyPoGen significantly outperforms state-of-the-art methods in generating policies for unseen target tasks without any demonstrations, achieving higher success rates and underscoring the potential of optimization-biased hypernetworks in advancing generalizable policy generation. Our code and data are available at: https://github.com/ReNginx/HyPoGen.

406. Learning Gain Map for Inverse Tone Mapping

链接：https://iclr.cc/virtual/2025/poster/30253 abstract： For a more compatible and consistent high dynamic range (HDR) viewing experience, a new image format with a double-layer structure has been developed recently, which incorporates an auxiliary Gain Map (GM) within a standard dynamic range (SDR) image for adaptive HDR display. This new format motivates us to introduce a new task termed Gain Map-based Inverse Tone Mapping (GM-ITM), which focuses on learning the corresponding GM of an SDR image instead of directly estimating its HDR counterpart, thereby enabling a more effective up-conversion by leveraging the advantages of GM. The main challenge in this task, however, is to accurately estimate regional intensity variation with the fluctuating peak value. To this end, we propose a dual-branch network named GMNet, consisting of a Local Contrast Restoration (LCR) branch and a Global Luminance Estimation (GLE) branch to capture pixel-wise and image-wise information for GM estimation. Moreover, to facilitate the future research of the GM-ITM task, we build both synthetic and real-world datasets for comprehensive evaluations: synthetic SDR-GM pairs are generated from existing HDR resources, and real-world SDR-GM pairs are captured by mobile devices. Extensive experiments on these datasets demonstrate the superiority of our proposed GMNet over existing HDR-related methods both quantitatively and qualitatively. The codes and datasets are available at https://github.com/qtlark/GMNet.

407. Causal Graph Transformer for Treatment Effect Estimation Under Unknown Interference

链接：https://iclr.cc/virtual/2025/poster/28858 abstract： Networked interference, also known as the peer effect in social science and spillover effect in economics, has drawn increasing interest across various domains. This phenomenon arises when a unit’s treatment and outcome are influenced by the actions of its peers, posing significant challenges to causal inference, particularly in treatment assignment and effect estimation in real applications, due to the violation of the SUTVA assumption. While extensive graph models have been developed to identify treatment effects, these models often rely on structural assumptions about networked interference, assuming it to be identical to the social network, which can lead to misspecification issues in real applications. To address these challenges, we propose an Interference-Agnostic Causal Graph Transformer (CauGramer), which aggregates peers information via L-order Graph Transformer and employs cross-attention to infer aggregation function for learning interference representations. By integrating confounder balancing and minimax moment constraints, CauGramer fully incorporates peer information, enabling robust treatment effect estimation. Extensive experiments on two widely-used benchmarks demonstrate the effectiveness and superiority of CauGramer. The code is available at https://github.com/anpwu/CauGramer.

408. Does SGD really happen in tiny subspaces?

链接：https://iclr.cc/virtual/2025/poster/27933 abstract： Understanding the training dynamics of deep neural networks is challenging due to their high-dimensional nature and intricate loss landscapes. Recent studies have revealed that, along the training trajectory, the gradient approximately aligns with a low-rank top eigenspace of the training loss Hessian, referred to as the dominant subspace. Given this alignment, this paper explores whether neural networks can be trained within the dominant subspace, which, if feasible, could lead to more efficient training methods. Our primary observation is that when the SGD update is projected onto the dominant subspace, the training loss does not decrease further.This suggests that the observed alignment between the gradient and the dominant subspace is spurious. Surprisingly, projecting out the dominant subspace proves to be just as effective as the original update, despite removing the majority of the original update component. We observe similar behavior across practical setups, including the large learning rate regime (also known as Edge of Stability), Sharpness-Aware Minimization, momentum, and adaptive optimizers. We discuss the main causes and implications of this spurious alignment, shedding light on the dynamics of neural network training.

409. The Superposition of Diffusion Models Using the Itô Density Estimator

链接：https://iclr.cc/virtual/2025/poster/31117 abstract： The Cambrian explosion of easily accessible pre-trained diffusion models suggests a demand for methods that combine multiple different pre-trained diffusion models without incurring the

significant computational burden of re-training a larger combined model. In this paper, we cast the problem of combining multiple pre-trained diffusion models at the generation stage under a novel proposed framework termed superposition. Theoretically, we derive superposition from rigorous first principles stemming from the celebrated continuity equation and design two novel algorithms tailor-made for combining diffusion models in SuperDiff. SuperDiff leverages a new scalable Itô density estimator for the log likelihood of the diffusion SDE which incurs no additional overhead compared to the well-known Hutchinson's estimator needed for divergence calculations. We demonstrate that SuperDiff is scalable to large pre-trained diffusion models as superposition is performed solely through composition during inference, and also enjoys painless implementation as it combines different pre-trained vector fields through an automated re-weighting scheme. Notably, we show that SuperDiff is efficient during inference time, and mimics traditional composition operators such as the logical OR and the logical AND. We empirically demonstrate the utility of using SuperDiff for generating more diverse images on CIFAR-10, more faithful prompt conditioned image editing using Stable Diffusion, as well as improved conditional molecule generation and unconditional de novo structure design of proteins. https://github.com/necludov/super-diffusion

410. CBQ: Cross-Block Quantization for Large Language Models

链接：https://iclr.cc/virtual/2025/poster/28924 abstract： Post-training quantization (PTQ) has played a pivotal role in compressing large language models (LLMs) at ultra-low costs. Although current PTQ methods have achieved promising results by addressing outliers and employing layer- or block-wise loss optimization techniques, they still suffer from significant performance degradation at ultra-low bits precision. To dissect this issue, we conducted an in-depth analysis of quantization errors specific to LLMs and surprisingly discovered that, unlike traditional sources of quantization errors, the growing number of model parameters, combined with the reduction in quantization bits, intensifies inter-layer and intra-layer dependencies, which severely impact quantization accuracy. This finding highlights a critical challenge in quantizing LLMs. To address this, we propose CBQ, a cross-block reconstruction-based PTQ method for LLMs. CBQ leverages a cross-block dependency to establish long-range dependencies across multiple blocks and integrates an adaptive LoRA-Rounding technique to manage intra-layer dependencies. To further enhance performance, CBQ incorporates a coarse-to-fine pre-processing mechanism for processing weights and activations. Extensive experiments show that CBQ achieves superior low-bit quantization (W4A4, W4A8, W2A16) and outperforms existing state-of-the-art methods across various LLMs and datasets. Notably, CBQ only takes 4.3 hours to quantize a weight-only quantization of a 4-bit LLAMA1-65B model, achieving a commendable trade off between performance and efficiency.

411. Amulet: ReAlignment During Test Time for Personalized Preference Adaptation of LLMs

链接：https://iclr.cc/virtual/2025/poster/28890 abstract： How to align large language models (LLMs) with user preferences from a static general dataset has been frequently studied. However, user preferences are usually personalized, changing, and diverse. This leads to the problem that the actual user preferences often do not coincide with those trained by the model developers in the practical use of LLMs. Since we cannot collect enough data and retrain for every demand, researching efficient real-time preference adaptation methods based on the backbone LLMs during test time is important. To this end, we introduce Amulet, a novel, training-free framework that formulates the decoding process of every token as a separate online learning problem with the guidance of simple user-provided prompts, thus enabling real-time optimization to satisfy users' personalized preferences. To reduce the computational cost brought by this optimization process for each token, we additionally provide a closed-form solution for each iteration step of the optimization process, thereby reducing the computational time cost to a negligible level. The detailed experimental results demonstrate that Amulet can achieve significant performance improvements in rich settings with combinations of different LLMs, datasets, and user preferences, while maintaining acceptable computational efficiency.

412. Weak-to-Strong Generalization Through the Data-Centric Lens

链接：https://iclr.cc/virtual/2025/poster/27959 abstract： The weak-to-strong generalization phenomenon is the driver for important machine learning applications including highly data-efficient learning and, most recently, performing superalignment. While decades of research have resulted in numerous algorithms that produce strong empirical performance, understanding what aspects of data enable weak-to-strong generalization has been understudied. We propose a simple data-centric mechanism that characterizes weak-to-strong generalization: the overlap density. Intuitively, generalization tracks the number of points that contain overlaps, i.e., both easy patterns (learnable by a weak model) and challenging patterns (only learnable by a stronger model), as with such points, weak predictions can be used to learn challenging patterns by stronger models. And, we provide a practical overlap detection algorithm to find overlap density from data. Finally, we provide an algorithm to learn, among multiple sources of data, which to query when seeking to maximize overlap density and thereby enhance weak-to-strong generalization. We provide a theoretical result showing that the generalization benefit is a function of the overlap density and a regret bound of our data selection algorithm. Empirically, we validate the mechanism and the overlap detection algorithm on a wide array of settings.

413. Magnetic Preference Optimization: Achieving Last-iterate Convergence for Language Model Alignment

链接：https://iclr.cc/virtual/2025/poster/29766 abstract： Self-play methods have demonstrated remarkable success in

enhancing model capabilities across various domains. In the context of Reinforcement Learning from Human Feedback (RLHF), self-play not only boosts Large Language Model (LLM) performance but also overcomes the limitations of traditional Bradley-Terry (BT) model assumptions by finding the Nash equilibrium (NE) of a preference-based, two-player constant-sum game. However, existing methods either guarantee only average-iterate convergence, incurring high storage and inference costs, or converge to the NE of a regularized game, failing to accurately reflect true human preferences. In this paper, we introduce Magnetic Preference Optimization (MPO), a novel approach capable of achieving last-iterate convergence to the NE of the original game, effectively overcoming the limitations of existing methods. Building upon Magnetic Mirror Descent (MMD), MPO attains a linear convergence rate, making it particularly suitable for fine-tuning LLMs. To ensure our algorithm is both theoretically sound and practically viable, we present a simple yet effective implementation that adapts the theoretical insights to the RLHF setting. Empirical results demonstrate that MPO can significantly enhance the performance of LLMs, highlighting the potential of self-play methods in alignment.

414. Implicit In-context Learning

链接：https://iclr.cc/virtual/2025/poster/30298 abstract： In-context Learning (ICL) empowers large language models (LLMs) to swiftly adapt to unseen tasks at inference-time by prefixing a few demonstration examples before queries. Despite its versatility, ICL incurs substantial computational and memory overheads compared to zero-shot learning and is sensitive to the selection and order of demonstration examples. In this work, we introduce \textbf{Implicit In-context Learning} (I2CL), an innovative paradigm that reduces the inference cost of ICL to that of zero-shot learning with minimal information loss. I2CL operates by first generating a condensed vector representation, namely a context vector, extracted from the demonstration examples. It then conducts an inference-time intervention through injecting a linear combination of the context vector and query activations back into the model's residual streams. Empirical evaluation on nine real-world tasks across three model architectures demonstrates that I2CL achieves few-shot level performance at zero-shot inference cost, and it exhibits robustness against variations in demonstration examples. Furthermore, I2CL facilitates a novel representation of ``task-ids", enhancing task similarity detection and fostering effective transfer learning. We also perform a comprehensive analysis and ablation study on I2CL, offering deeper insights into its internal mechanisms. Code is available at https://github.com/LzVv123456/I2CL.

415. Understanding the Stability-based Generalization of Personalized Federated Learning

链接：https://iclr.cc/virtual/2025/poster/27649 abstract： Despite great achievements in algorithm design for Personalized Federated Learning (PFL), research on the theoretical analysis of generalization is still in its early stages. Some theoretical results have investigated the generalization performance of personalized models under the problem setting and hypothesis in convex conditions, which can not reflect the real iteration performance during non-convex training. To further understand the real performance from a generalization perspective, we propose the first algorithm-dependent generalization analysis with uniform stability for the typical PFL method, Partial Model Personalization, on smooth and non-convex objectives. Specifically, we decompose the generalization errors into aggregation errors and fine-tuning errors, then creatively establish a generalization analysis framework corresponding to the gradient estimation process of the personalized training. This framework builds up the bridge among PFL, FL and Pure Local Training for personalized aims in heterogeneous scenarios, which clearly demonstrates the effectiveness of PFL from the generalization perspective. Moreover, we demonstrate the impact of trivial factors like learning steps, stepsizes and communication topologies and obtain the excess risk analysis with optimization errors for PFL. Promising experiments on CIFAR datasets also corroborate our theoretical insights. Our code can be seen in https://github.com/YingqiLiu1999/Understanding-the-Stability-based-Generalization-of-Personalized-Federated-Learning.

416. Data-centric Prediction Explanation via Kernelized Stein Discrepancy

链接：https://iclr.cc/virtual/2025/poster/30035 abstract： Existing example-based prediction explanation methods often bridge test and training data points through the model's parameters or latent representations. While these methods offer clues to the causes of model predictions, they often exhibit innate shortcomings, such as incurring significant computational overhead or producing coarse-grained explanations. This paper presents a Highly-precise and Data-centric Explanation (HD-Explain) prediction explanation method that exploits properties of Kernelized Stein Discrepancy (KSD). Specifically, the KSD uniquely defines a parameterized kernel function for a trained model that encodes model-dependent data correlation. By leveraging the kernel function, one can identify training samples that provide the best predictive support to a test point efficiently. We conducted thorough analyses and experiments across multiple classification domains, where we show that HD-Explain outperforms existing methods from various aspects, including 1) preciseness (fine-grained explanation), 2) consistency, and 3) computation efficiency, leading to a surprisingly simple, effective, and robust prediction explanation solution.

417. GenVP: Generating Visual Puzzles with Contrastive Hierarchical VAEs

链接：https://iclr.cc/virtual/2025/poster/29743 abstract： Raven's Progressive Matrices (RPMs) is an established benchmark to examinethe ability to perform high-level abstract visual reasoning (AVR). Despite the current success of algorithms that solve this task, humans can generalize beyond a given puzzle and create new puzzles given a set of rules, whereas machines remain locked in solving a fixed puzzle from a curated choice list. We propose Generative Visual Puzzles (GenVP), a framework to model the entire RPM generation process, a substantially more challenging task. Our model's capability spans from generating multiple solutions for one specific problem prompt to creating complete new puzzles out of the desired set of rules. Experiments on five different datasets indicate that GenVP achieves state-of-the-art (SOTA) performance both in puzzle-solving accuracy and

out-of-distribution (OOD) generalization in 22 outof 24 OOD scenarios. Further, compared to SOTA generative approaches, which struggle to solve RPMs when the feasible solution space increases, GenVP efficiently generalizes to these challenging scenarios. Moreover, our model demonstrates the ability to produce a wide range of complete RPMs given a set of abstract rules by effectively capturing the relationships between abstract rules and visual object properties.

418. MotionDreamer: One-to-Many Motion Synthesis with Localized Generative Masked Transformer

链接：https://iclr.cc/virtual/2025/poster/29015 abstract： Generative masked transformer have demonstrated remarkable success across various content generation tasks, primarily due to their ability to effectively model large-scale dataset distributions with high consistency. However, in the animation domain, large datasets are not always available. Applying generative masked modeling to generate diverse instances from a single MoCap reference may lead to overfitting, a challenge that remains unexplored. In this work, we present MotionDreamer, a localized masked modeling paradigm designed to learn motion internal patterns from a given motion with arbitrary topology and duration. By embedding the given motion into quantized tokens with a novel distribution regularization method, MotionDreamer constructs a robust and informative codebook for local motion patterns. Moreover, a sliding window local attention is introduced in our masked transformer, enabling the generation of natural yet diverse animations that closely resemble the reference motion patterns. As demonstrated through comprehensive experiments, MotionDreamer outperforms the state-of-the-art methods that are typically GAN or Diffusion-based in both faithfulness and diversity. Thanks to the consistency and robustness of quantization-based approach, MotionDreamer can also effectively perform downstream tasks such as temporal motion editing, crowd motion synthesis, and beat-aligned dance generation, all using a single reference motion. Our implementation, learned models and results are to be made publicly available upon paper acceptance.

419. Standard Gaussian Process is All You Need for High-Dimensional Bayesian Optimization

链接：https://iclr.cc/virtual/2025/poster/28574 abstract： A long-standing belief holds that Bayesian Optimization (BO) with standard Gaussian processes (GP) --- referred to as standard BO --- underperforms in high-dimensional optimization problems. While this belief seems plausible, it lacks both robust empirical evidence and theoretical justification. To address this gap, we present a systematic investigation. First, through a comprehensive evaluation across twelve benchmarks, we found that while the popular Square Exponential (SE) kernel often leads to poor performance, using Mat\'ern kernels enables standard BO to consistently achieve top-tier results, frequently surpassing methods specifically designed for high-dimensional optimization. Second, our theoretical analysis reveals that the SE kernel’s failure primarily stems from improper initialization of the length-scale parameters, which are commonly used in practice but can cause gradient vanishing in training. We provide a probabilistic bound to characterize this issue, showing that Mat\'ern kernels are less susceptible and can robustly handle much higher dimensions. Third, we propose a simple robust initialization strategy that dramatically improves the performance of the SE kernel, bringing it close to state-of-the-art methods, without requiring additional priors or regularization. We prove another probabilistic bound that demonstrates how the gradient vanishing issue can be effectively mitigated with our method. Our findings advocate for a re-evaluation of standard BO’s potential in high-dimensional settings.

420. Scaling Speech-Text Pre-training with Synthetic Interleaved Data

链接：https://iclr.cc/virtual/2025/poster/32109 abstract： Speech language models (SpeechLMs) accept speech input and produce speech output, allowing for more natural human-computer interaction compared to text-based large language models (LLMs).Traditional approaches for developing SpeechLMs are constrained by the limited availability of unsupervised speech data and parallel speech-text data, which are significantly less abundant compared to text pre-training data, thereby limiting their scalability as LLMs.We propose a novel approach to scaling speech-text pre-training by leveraging large-scale synthetic interleaved data derived from text corpora, eliminating the need for parallel speech-text datasets.Our method efficiently constructs speech-text interleaved data by sampling text spans from existing text corpora and synthesizing corresponding speech spans using a text-to-token model, bypassing the need to generate actual speech.We also employ a supervised speech tokenizer derived from an automatic speech recognition (ASR) model by incorporating a vector-quantized bottleneck into the encoder. This supervised training approach results in discrete speech tokens with strong semantic preservation even at lower sampling rates (e.g. 12.5Hz), while still maintaining speech reconstruction quality.Starting from a pre-trained language model and scaling our pre-training to 1 trillion tokens (with 600B synthetic interleaved speech-text data), we achieve state-of-the-art performance in both speech language modeling and spoken question answering, improving performance on spoken questions tasks from the previous SOTA of 13\% (Moshi) to 31\%.We further demonstrate that by fine-tuning the pre-trained model with speech dialogue data, we can develop an end-to-end spoken chatbot that achieves competitive performance comparable to existing baselines in both conversational abilities and speech quality, even operating exclusively in the speech domain.

421. Global Convergence of Policy Gradient in Average Reward MDPs

链接：https://iclr.cc/virtual/2025/poster/31141 abstract： We present the first comprehensive finite-time global convergence analysis of policy gradient for infinite horizon average reward Markov decision processes (MDPs). Specifically, we focus on ergodic tabular MDPs with finite state and action spaces. Our analysis shows that the policy gradient iterates converge to the optimal policy at a sublinear rate of $O(\frac{1}{T})$, where T represents the number of iterations. Performance bounds for

discounted reward MDPs cannot be easily extended to average reward MDPs as the bounds grow proportional to the fifth power of the effective horizon. Recent work on such extensions makes a smoothness assumption that has not been verified. Thus, our primary contribution is in providing the first complete proof that the policy gradient algorithm converges globally for average-reward MDPs, without such an assumption. We also obtain the corresponding finite-time performance guarantees. In contrast to the existing discounted reward performance bounds, our performance bounds have an explicit dependence on constants that capture the complexity of the underlying MDP. Motivated by this observation, we reexamine and improve the existing performance bounds for discounted reward MDPs. We also present simulations that empirically validate the result.

422. Learning-Augmented Frequent Directions

链接：https://iclr.cc/virtual/2025/poster/29356 abstract： An influential paper of Hsu et al. (ICLR'19) introduced the study of learning-augmented streaming algorithms in the context of frequency estimation. A fundamental problem in the streaming literature, the goal of frequency estimation is to approximate the number of occurrences of items appearing in a long stream of data using only a small amount of memory. Hsu et al. develop a natural framework to combine the worst-case guarantees of popular solutions such as CountMin and CountSketch with learned predictions of high frequency elements. They demonstrate that learning the underlying structure of data can be used to yield better streaming algorithms, both in theory and practice.We simplify and generalize past work on learning-augmented frequency estimation. Our first contribution is a learning-augmented variant of the Misra-Gries algorithm which improves upon the error of learned CountMin and learned CountSketch and achieves the state-of-the-art performance of randomized algorithms (Aamand et al., NeurIPS'23) with a simpler, deterministic algorithm. Our second contribution is to adapt learning-augmentation to a high-dimensional generalization of frequency estimation corresponding to finding important directions (top singular vectors) of a matrix given its rows one-by-one in a stream. We analyze a learning-augmented variant of the Frequent Directions algorithm, extending the theoretical and empirical understanding of learned predictions to matrix streaming.

423. VisualAgentBench: Towards Large Multimodal Models as Visual Foundation Agents

链接：https://iclr.cc/virtual/2025/poster/31108 abstract： Large Multimodal Models (LMMs) have ushered in a new era in artificial intelligence, merging capabilities in both language and vision to form highly capable \textbf{Visual Foundation Agents} that are postulated to excel across a myriad of tasks. However, existing benchmarks fail to sufficiently challenge or showcase the full potential of LMMs as visual foundation agents in complex, real-world environments. To address this gap, we introduce VisualAgentBench (VAB), a comprehensive and unified benchmark specifically designed to train and evaluate LMMs as visual foundation agents across diverse scenarios in one standard setting, including Embodied, Graphical User Interface, and Visual Design, with tasks formulated to probe the depth of LMMs' understanding and interaction capabilities. Through rigorous testing across 9 proprietary LMM APIs and 9 open models (18 in total), we demonstrate the considerable yet still developing visual agent capabilities of these models. Additionally, VAB explores the synthesizing of visual agent trajectory data through hybrid methods including Program-based Solvers, LMM Agent Bootstrapping, and Human Demonstrations, offering insights into obstacles, solutions, and trade-offs one may meet in developing open LMM agents. Our work not only aims to benchmark existing models but also provides an instrumental playground for future development into visual foundation agents. Code, train, and test data are available at \url{https://github.com/THUDM/VisualAgentBench}.

424. SV4D: Dynamic 3D Content Generation with Multi-Frame and Multi-View Consistency

链接：https://iclr.cc/virtual/2025/poster/28061 abstract： We present Stable Video 4D (SV4D) — a latent video diffusion model for multi-frame and multi-view consistent dynamic 3D content generation. Unlike previous methods that rely on separately trained generative models for video generation and novel view synthesis, we design a unified diffusion model to generate novel view videos of dynamic 3D objects. Specifically, given a monocular reference video, SV4D generates novel views for each video frame that are temporally consistent. We then use the generated novel view videos to optimize an implicit 4D representation (dynamic NeRF) efficiently, without the need for cumbersome SDS-based optimization used in most prior works. To train our unified novel view video generation model, we curate a dynamic 3D object dataset from the existing Objaverse dataset. Extensive experimental results on multiple datasets and user studies demonstrate SV4D's state-of-the-art performance on novel-view video synthesis as well as 4D generation compared to prior works. Project page: https://sv4d.github.io.

425. Pareto Low-Rank Adapters: Efficient Multi-Task Learning with Preferences

链接：https://iclr.cc/virtual/2025/poster/28684 abstract： Multi-task trade-offs in machine learning can be addressed via Pareto Front Learning (PFL) methods that parameterize the Pareto Front (PF) with a single model. PFL permits to select the desired operational point during inference, contrary to traditional Multi-Task Learning (MTL) that optimizes for a single trade-off decided prior to training. However, recent PFL methodologies suffer from limited scalability, slow convergence, and excessive memory requirements, while exhibiting inconsistent mappings from preference to objective space. We introduce PaLoRA, a novel parameter-efficient method that addresses these limitations in two ways. First, we augment any neural network architecture with task-specific low-rank adapters and continuously parameterize the Pareto Front in their convex hull. Our approach steers the

original model and the adapters towards learning general and task-specific features, respectively. Second, we propose a deterministic sampling schedule of preference vectors that reinforces this division of labor, enabling faster convergence and strengthening the validity of the mapping from preference to objective space throughout training. Our experiments show that PaLoRA outperforms state-of-the-art MTL and PFL baselines across various datasets, scales to large networks, reducing the memory overhead $23.8-31.7$ times compared with competing PFL baselines in scene understanding benchmarks.

426. MA-RLHF: Reinforcement Learning from Human Feedback with Macro Actions

链接：https://iclr.cc/virtual/2025/poster/29359 abstract： Reinforcement learning from human feedback (RLHF) has demonstrated effectiveness in aligning large language models (LLMs) with human preferences. However, token-level RLHF suffers from the credit assignment problem over long sequences, where delayed rewards make it challenging for the model to discern which actions contributed to preferred outcomes. This hinders learning efficiency and slows convergence.In this paper, we propose MA-RLHF, a simple yet effective RLHF framework that incorporates macro actions --- sequences of tokens or higher-level language constructs --- into the learning process. By operating at higher level of abstraction, our approach reduces the temporal distance between actions and rewards, facilitating faster and more accurate credit assignment. This results in more stable policy gradient estimates and enhances learning efficiency within each episode, all without increasing computational complexity during training or inference. We validate our approach through extensive experiments across various model sizes and tasks, including text summarization, dialogue generation, question answering, and program synthesis. Our method achieves substantial performance improvements over standard RLHF, with performance gains of up to 30\% in text summarization and code generation, 18\% in dialogue, and 8\% in question answering tasks. Notably, our approach reaches parity with vanilla RLHF $1.7 \sim 2$ times faster in terms of training time and continues to outperform it with further training. We make our code and data publicly available at \url{https://github.com/ernie-research/MA-RLHF}.

427. Decoupling Layout from Glyph in Online Chinese Handwriting Generation

链接：https://iclr.cc/virtual/2025/poster/30447 abstract： Text plays a crucial role in the transmission of human civilization, and teaching machines to generate online handwritten text in various styles presents an interesting and significant challenge. However, most prior work has concentrated on generating individual Chinese fonts, leaving complete text line generation largely unexplored. In this paper, we identify that text lines can naturally be divided into two components: layout and glyphs. Based on this division, we designed a text line layout generator coupled with a diffusion-based stylized font synthesizer to address this challenge hierarchically. More concretely, the layout generator performs in-context-like learning based on the text content and the provided style references to generate positions for each glyph autoregressively. Meanwhile, the font synthesizer which consists of a character embedding dictionary, a multi-scale calligraphy style encoder and a 1D U-Net based diffusion denoiser will generate each font on its position while imitating the calligraphy style extracted from the given style references. Qualitative and quantitative experiments on the CASIA-OLHWDB demonstrate that our method is capable of generating structurally correct and indistinguishable imitation samples.

428. Structuring Benchmark into Knowledge Graphs to Assist Large Language Models in Retrieving and Designing Models

链接：https://iclr.cc/virtual/2025/poster/31033 abstract： In recent years, the design and transfer of neural network models have been widely studied due to their exceptional performance and capabilities. However, the complex nature of datasets and the vast architecture space pose significant challenges for both manual and automated algorithms in creating high-performance models. Inspired by researchers who design, train, and document the performance of various models across different datasets, this paper introduces a novel schema that transforms the benchmark data into a Knowledge Benchmark Graph (KBG), which primarily stores the facts in the form of performance(data, model). Constructing the KBG facilitates the structured storage of design knowledge, aiding subsequent model design and transfer. However, it is a non-trivial task to retrieve or design suitable neural networks based on the KBG, as real-world data are often off the records. To tackle this challenge, we propose transferring existing models stored in KBG by establishing correlations between unseen and previously seen datasets. Given that measuring dataset similarity is a complex and open-ended issue, we explore the potential for evaluating the correctness of the similarity function. Then, we further integrate the KBG with Large Language Models (LLMs), assisting LLMs to think and retrieve existing model knowledge in a manner akin to humans when designing or transferring models. We demonstrate our method specifically in the context of Graph Neural Network (GNN) architecture design, constructing a KBG (with 26,206 models, 211,669 performance records, and 2,540,064 facts) and validating the effectiveness of leveraging the KBG to promote GNN architecture design.

429. A Quantum Circuit-Based Compression Perspective for Parameter-Efficient Learning

链接：https://iclr.cc/virtual/2025/poster/29122 abstract： Quantum-centric supercomputing presents a compelling framework for large-scale hybrid quantum-classical tasks. Although quantum machine learning (QML) offers theoretical benefits in various applications, challenges such as large-size data encoding in the input stage and the reliance on quantum resources in the

inference stage limit its practicality for tasks like fine-tuning large language models (LLMs). Quantum parameter generation, a novel approach of QML, addresses these limitations by using quantum neural networks (QNNs) to generate classical model weights (parameters) exclusively during training, thereby decoupling inference from quantum hardware. In this work, we introduce Quantum Parameter Adaptation (QPA) in the framework of quantum parameter generation, which integrates QNNs with a classical multi-layer perceptron mapping model to generate parameters for fine-tuning methods. Using Gemma-2 and GPT-2 as case studies, QPA demonstrates significant parameter reduction for parameter-efficient fine-tuning methods, such as Low-Rank Adaptation (LoRA), while maintaining comparable or improved performance in text generation tasks. Specifically, QPA reduces the number of parameters to 52.06% of the original LoRA for GPT-2 with a slight performance gain of 0.75%, and to 16.84% for Gemma-2, with a marginal performance improvement of 0.07%. These results highlight QPA's ability to achieve efficient parameter reduction without sacrificing performance in the quantum parameter generation framework. This work showcases the potential of quantum-enhanced parameter reduction, offering a scalable quantum-classical solution for fine-tuning LLMs while preserving the feasibility of inference on classical hardware.

430. Causal Information Prioritization for Efficient Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/28418 abstract： Current Reinforcement Learning (RL) methods often suffer from sample-inefficiency, resulting from blind exploration strategies that neglect causal relationships among states, actions, and rewards. Although recent causal approaches aim to address this problem, they lack grounded modeling of reward-guided causal understanding of states and actions for goal-orientation, thus impairing learning efficiency. To tackle this issue, we propose a novel method named Causal Information Prioritization (CIP) that improves sample efficiency by leveraging factored MDPs to infer causal relationships between different dimensions of states and actions with respect to rewards, enabling the prioritization of causal information. Specifically, CIP identifies and leverages causal relationships between states and rewards to execute counterfactual data augmentation to prioritize high-impact state features under the causal understanding of the environments. Moreover, CIP integrates a causality-aware empowerment learning objective, which significantly enhances the agent's execution of reward-guided actions for more efficient exploration in complex environments. To fully assess the effectiveness of CIP, we conduct extensive experiments across 39 tasks in 5 diverse continuous control environments, encompassing both locomotion and manipulation skills learning with pixel-based and sparse reward settings. Experimental results demonstrate that CIP consistently outperforms existing RL methods across a wide range of scenarios.

431. SRSA: Skill Retrieval and Adaptation for Robotic Assembly Tasks

链接：https://iclr.cc/virtual/2025/poster/29656 abstract： Enabling robots to learn novel tasks in a data-efficient manner is a long-standing challenge. Common strategies involve carefully leveraging prior experiences, especially transition data collected on related tasks. Although much progress has been made for general pick-and-place manipulation, far fewer studies have investigated contact-rich assembly tasks, where precise control is essential. We introduce SRSA} (Skill Retrieval and Skill Adaptation), a novel framework designed to address this problem by utilizing a pre-existing skill library containing policies for diverse assembly tasks. The challenge lies in identifying which skill from the library is most relevant for fine-tuning on a new task. Our key hypothesis is that skills showing higher zero-shot success rates on a new task are better suited for rapid and effective fine-tuning on that task. To this end, we propose to predict the transfer success for all skills in the skill library on a novel task, and then use this prediction to guide the skill retrieval process. We establish a framework that jointly captures features of object geometry, physical dynamics, and expert actions to represent the tasks, allowing us to efficiently learn the transfer success predictor. Extensive experiments demonstrate that SRSA significantly outperforms the leading baseline. When retrieving and fine-tuning skills on unseen tasks, SRSA achieves a 19% relative improvement in success rate, exhibits 2.6x lower standard deviation across random seeds, and requires 2.4x fewer transition samples to reach a satisfactory success rate, compared to the baseline. In a continual learning setup, SRSA efficiently learns policies for new tasks and incorporates them into the skill library, enhancing future policy learning. Furthermore, policies trained with SRSA in simulation achieve a 90% mean success rate when deployed in the real world. Please visit our project webpage https://srsa2024.github.io/.

432. Robotouille: An Asynchronous Planning Benchmark for LLM Agents

链接：https://iclr.cc/virtual/2025/poster/29809 abstract： Effective asynchronous planning, or the ability to efficiently reason and plan over states and actions that must happen in parallel or sequentially, is essential for agents that must account for time delays, reason over diverse long-horizon tasks, and collaborate with other agents. While large language model (LLM) agents show promise in high-level task planning, current benchmarks focus primarily on short-horizon tasks and do not evaluate such asynchronous planning capabilities. We introduce Robotouille, a challenging benchmark environment designed to test LLM agents' ability to handle long-horizon asynchronous scenarios. Our synchronous and asynchronous datasets capture increasingly complex planning challenges that go beyond existing benchmarks, requiring agents to manage over-lapping tasks and interruptions Our results show that ReAct (gpt-4o) achieves 47% on synchronous tasks but only 11% on asynchronous tasks, highlighting significant room for improvement. We further analyze failure modes, demonstrating the need for LLM agents to better incorporate long-horizon feedback and self-audit their reasoning during task execution.

433. Learning local equivariant representations for quantum operators

链接：https://iclr.cc/virtual/2025/poster/28559 abstract： Predicting quantum operator matrices such as Hamiltonian, overlap, and density matrices in the density functional theory (DFT) framework is crucial for material science. Current methods often focus on individual operators and struggle with efficiency and scalability for large systems. Here we introduce a novel deep

learning model, SLEM (strictly localized equivariant message-passing), for predicting multiple quantum operators that achieves state-of-the-art accuracy while dramatically improving computational efficiency. SLEM's key innovation is its strict locality-based design for equivariant representations of quantum tensors while preserving physical symmetries. This enables complex many-body dependency without expanding the effective receptive field, leading to superior data efficiency and transferability. Using an innovative SO(2) convolution and invariant overlap parameterization, SLEM reduces the computational complexity of high-order tensor products and is, therefore, capable of handling systems requiring the f and g orbitals in their basis sets. We demonstrate SLEM's capabilities across diverse 2D and 3D materials, achieving high accuracy even with limited training data. SLEM's design facilitates efficient parallelization, potentially extending DFT simulations to systems with device-level sizes, opening new possibilities for large-scale quantum simulations and high-throughput materials discovery.

434. Towards Auto-Regressive Next-Token Prediction: In-context Learning Emerges from Generalization

链接：https://iclr.cc/virtual/2025/poster/28825 abstract： Large language models (LLMs) have demonstrated remarkable in-context learning (ICL) abilities. However, existing theoretical analysis of ICL primarily exhibits two limitations: \textbf{(a) Limited \textit{i.i.d.} Setting.} Most studies focus on supervised function learning tasks where prompts are constructed with \textit{i.i.d.} input-label pairs. This \textit{i.i.d.} assumption diverges significantly from real language learning scenarios where prompt tokens are interdependent. \textbf{(b) Lack of Emergence Explanation.} Most literature answers \textbf{\textit{what}} ICL does from an implicit optimization perspective but falls short in elucidating \textbf{\textit{how}} ICL emerges and the impact of pre-training phase on ICL. In our paper, to extend (a), we adopt a more practical paradigm, \textbf{\textit{auto-regressive next-token prediction (AR-NTP)}}, which closely aligns with the actual training of language models. Specifically, within AR-NTP, we emphasize prompt token-dependency, which involves predicting each subsequent token based on the preceding sequence. To address (b), we formalize a systematic pre-training and ICL framework, highlighting the layer-wise structure of sequences and topics, alongside a two-level expectation. In conclusion, we present data-dependent, topic-dependent and optimization-dependent PAC-Bayesian generalization bounds for pre-trained LLMs, investigating that \textbf{\textit{ICL emerges from the generalization of sequences and topics}}. Our theory is supported by experiments on numerical linear dynamic systems, synthetic GINC and real-world language datasets.

435. TTVD: Towards a Geometric Framework for Test-Time Adaptation Based on Voronoi Diagram

链接：https://iclr.cc/virtual/2025/poster/30927 abstract： Deep learning models often struggle with generalization when deploying on real-world data, due to the common distributional shift to the training data. Test-time adaptation (TTA) is an emerging scheme used at inference time to address this issue. In TTA, models are adapted online at the same time when making predictions to test data. Neighbor-based approaches have gained attention recently, where prototype embeddings provide location information to alleviate the feature shift between training and testing data. However, due to their inherit limitation of simplicity, they often struggle to learn useful patterns and encounter performance degradation. To confront this challenge, we study the TTA problem from a geometric point of view. We first reveal that the underlying structure of neighbor-based methods aligns with the Voronoi Diagram, a classical computational geometry model for space partitioning. Building on this observation, we propose the Test-Time adjustment by Voronoi Diagram guidance (TTVD), a novel framework that leverages the benefits of this geometric property. Specifically, we explore two key structures: 1) Cluster-induced Voronoi Diagram (CIVD): This integrates the joint contribution of self-supervision and entropy-based methods to provide richer information. 2) Power Diagram (PD): A generalized version of the Voronoi Diagram that refines partitions by assigning weights to each Voronoi cell. Our experiments under rigid, peer-reviewed settings on CIFAR-10-C, CIFAR-100-C, ImageNet-C, and ImageNet-R shows that TTVD achieves remarkable improvements compared to state-of-the-art methods. Moreover, extensive experimental results also explore the effects of batch size and class imbalance, which are two scenarios commonly encountered in real-world applications. These analyses further validate the robustness and adaptability of our proposed framework.

436. Towards Empowerment Gain through Causal Structure Learning in Model-Based Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/27900 abstract： In Model-Based Reinforcement Learning (MBRL), incorporating causal structures into dynamics models provides agents with a structured understanding of the environments, enabling efficient decision. Empowerment as an intrinsic motivation enhances the ability of agents to actively control their environments by maximizing the mutual information between future states and actions. We posit that empowerment coupled with causal understanding can improve controllability, while enhanced empowerment gain can further facilitate causal reasoning in MBRL. To improve learning efficiency and controllability, we propose a novel framework, Empowerment through Causal Learning (ECL), where an agent with the awareness of causal dynamics models achieves empowerment-driven exploration and optimizes its causal structure for task learning. Specifically, ECL operates by first training a causal dynamics model of the environment based on collected data. We then maximize empowerment under the causal structure for exploration, simultaneously using data gathered through exploration to update causal dynamics model to be more controllable than dense dynamics model without causal structure. In downstream task learning, an intrinsic curiosity reward is included to balance the causality, mitigating overfitting. Importantly, ECL is method-agnostic and is capable of integrating various causal discovery methods. We evaluate ECL combined with 3 causal discovery methods across 6 environments including pixel-based tasks, demonstrating its superior performance compared to other causal MBRL methods, in terms of causal discovery, sample efficiency, and

asymptotic performance.

437. LiNeS: Post-training Layer Scaling Prevents Forgetting and Enhances Model Merging

链接：https://iclr.cc/virtual/2025/poster/30130 abstract： Fine-tuning pre-trained models has become the standard approach to endow them with specialized knowledge, but it poses fundamental challenges. In particular, (i) fine-tuning often leads to catastrophic forgetting, where improvements on a target domain degrade generalization on other tasks, and (ii) merging fine-tuned checkpoints from disparate tasks can lead to significant performance loss. To address these challenges, we introduce LiNeS, Layer-increasing Network Scaling, a post-training editing technique designed to preserve pre-trained generalization while enhancing fine-tuned task performance. LiNeS scales parameter updates linearly based on their layer depth within the network, maintaining shallow layers close to their pre-trained values to preserve general features while allowing deeper layers to retain task-specific representations. In multi-task model merging scenarios, layer-wise scaling of merged parameters reduces negative task interference. LiNeS demonstrates significant improvements in both single-task and multi-task settings across various benchmarks in vision and natural language processing. It mitigates forgetting, enhances out-of-distribution generalization, integrates seamlessly with existing multi-task model merging baselines improving their performance across benchmarks and model sizes, and can boost generalization when merging LLM policies aligned with different rewards via RLHF. Our method is simple to implement, computationally efficient and complementary to many existing techniques. Our source code is available at github.com/wang-kee/LiNeS.

438. Accelerating Training with Neuron Interaction and Nowcasting Networks

链接：https://iclr.cc/virtual/2025/poster/29046 abstract： Neural network training can be accelerated when a learnable update rule is used in lieu of classic adaptive optimizers (e.g. Adam). However, learnable update rules can be costly and unstable to train and use. Recently, Jang et al. (2023) proposed a simpler approach to accelerate training based on weight nowcaster networks (WNNs). In their approach, Adam is used for most of the optimization steps and periodically, only every few steps, a WNN nowcasts (predicts near future) parameters. We improve WNNs by proposing neuron interaction and nowcasting (NiNo) networks. In contrast to WNNs, NiNo leverages neuron connectivity and graph neural networks to more accurately nowcast parameters. We further show that in some networks, such as Transformers, modeling neuron connectivity accurately is challenging. We address this and other limitations, which allows NiNo to accelerate Adam training by up to 50% in vision and language tasks.

439. Phidias: A Generative Model for Creating 3D Content from Text, Image, and 3D Conditions with Reference-Augmented Diffusion

链接：https://iclr.cc/virtual/2025/poster/29554 abstract： Generative 3D modeling has made significant advances recently, but it remains constrained by its inherently ill-posed nature, leading to challenges in quality and controllability. Inspired by the real-world workflow that designers typically refer to existing 3D models when creating new ones, we propose Phidias, a novel generative model that uses diffusion for reference-augmented 3D generation. Given an image, our method leverages a retrieved or user-provided 3D reference model to guide the generation process, thereby enhancing the generation quality, generalization ability, and controllability. Phidias integrates three key components: 1) meta-ControlNet to dynamically modulate the conditioning strength, 2) dynamic reference routing to mitigate misalignment between the input image and 3D reference, and 3) self-reference augmentations to enable self-supervised training with a progressive curriculum. Collectively, these designs result in significant generative improvements over existing methods. Phidias forms a unified framework for 3D generation using text, image, and 3D conditions, offering versatile applications.

440. SiReRAG: Indexing Similar and Related Information for Multihop Reasoning

链接：https://iclr.cc/virtual/2025/poster/27699 abstract： Indexing is an important step towards strong performance in retrieval-augmented generation (RAG) systems. However, existing methods organize data based on either semantic similarity (similarity) or related information (relatedness), but do not cover both perspectives comprehensively. Our analysis reveals that modeling only one perspective results in insufficient knowledge synthesis, leading to suboptimal performance on complex tasks requiring multihop reasoning. In this paper, we propose SiReRAG, a novel RAG indexing approach that explicitly considers both similar and related information. On the similarity side, we follow existing work and explore some variances to construct a similarity tree based on recursive summarization. On the relatedness side, SiReRAG extracts propositions and entities from texts, groups propositions via shared entities, and generates recursive summaries to construct a relatedness tree. We index and flatten both similarity and relatedness trees into a unified retrieval pool. Our experiments demonstrate that SiReRAG consistently outperforms state-of-the-art indexing methods on three multihop datasets (MuSiQue, 2WikiMultiHopQA, and HotpotQA), with an average 1.9% improvement in F1 scores. As a reasonably efficient solution, SiReRAG enhances existing reranking methods significantly, with up to 7.8% improvement in average F1 scores. Our code is available at https://github.com/SalesforceAIResearch/SiReRAG.

441. DynaPrompt: Dynamic Test-Time Prompt Tuning

链接：https://iclr.cc/virtual/2025/poster/30415 abstract： Test-time prompt tuning enhances zero-shot generalization of vision-language models but tends to ignore the relatedness among test samples during inference. Online test-time prompt tuning provides a simple way to leverage the information in previous test samples, albeit with the risk of prompt collapse due to error accumulation. To enhance test-time prompt tuning, we propose DynaPrompt, short for dynamic test-time prompt tuning, exploiting relevant data distribution information while reducing error accumulation. Built on an online prompt buffer, DynaPrompt adaptively selects and optimizes the relevant prompts for each test sample during tuning. Specifically, we introduce a dynamic prompt selection strategy based on two metrics: prediction entropy and probability difference. For unseen test data information, we develop dynamic prompt appending, which allows the buffer to append new prompts and delete the inactive ones. By doing so, the prompts are optimized to exploit beneficial information on specific test data, while alleviating error accumulation. Experiments on fourteen datasets demonstrate the effectiveness of dynamic test-time prompt tuning.

442. MMDT: Decoding the Trustworthiness and Safety of Multimodal Foundation Models

链接：https://iclr.cc/virtual/2025/poster/28256 abstract： Multimodal foundation models (MMFMs) play a crucial role in various applications, including autonomous driving, healthcare, and virtual assistants. However, several studies have revealed vulnerabilities in these models, such as generating unsafe content by text-to-image models. Existing benchmarks on multimodal models either predominantly assess the helpfulness of these models, or only focus on limited perspectives such as fairness and privacy. In this paper, we present the first unified platform, MMDT (Multimodal DecodingTrust), designed to provide a comprehensive safety and trustworthiness evaluation for MMFMs. Our platform assesses models from multiple perspectives, including safety, hallucination, fairness/bias, privacy, adversarial robustness, and out-of-distribution (OOD) generalization. We have designed various evaluation scenarios and red teaming algorithms under different tasks for each perspective to generate challenging data, forming a high-quality benchmark. We evaluate a range of multimodal models using MMDT, and our findings reveal a series of vulnerabilities and areas for improvement across these perspectives. This work introduces the first comprehensive and unique safety and trustworthiness evaluation platform for MMFMs, paving the way for developing safer and more reliable MMFMs and systems. Our platform and benchmark are available at https://mmdecodingtrust.github.io/.

443. GS-LiDAR: Generating Realistic LiDAR Point Clouds with Panoramic Gaussian Splatting

链接：https://iclr.cc/virtual/2025/poster/29653 abstract： LiDAR novel view synthesis (NVS) has emerged as a novel task within LiDAR simulation, offering valuable simulated point cloud data from novel viewpoints to aid in autonomous driving systems. However, existing LiDAR NVS methods typically rely on neural radiance fields (NeRF) as their 3D representation, which incurs significant computational costs in both training and rendering. Moreover, NeRF and its variants are designed for symmetrical scenes, making them ill-suited for driving scenarios. To address these challenges, we propose GS-LiDAR, a novel framework for generating realistic LiDAR point clouds with panoramic Gaussian splatting. Our approach employs 2D Gaussian primitives with periodic vibration properties, allowing for precise geometric reconstruction of both static and dynamic elements in driving scenarios. We further introduce a novel panoramic rendering technique with explicit ray-splat intersection, guided by panoramic LiDAR supervision. By incorporating intensity and ray-drop spherical harmonic (SH) coefficients into the Gaussian primitives, we enhance the realism of the rendered point clouds. Extensive experiments on KITTI-360 and nuScenes demonstrate the superiority of our method in terms of quantitative metrics, visual quality, as well as training and rendering efficiency.

444. Forgetting Transformer: Softmax Attention with a Forget Gate

链接：https://iclr.cc/virtual/2025/poster/28268 abstract： An essential component of modern recurrent sequence models is the forget gate. While Transformers do not have an explicit recurrent form, we show that a forget gate can be naturally incorporated into Transformers by down-weighting the unnormalized attention scores in a data-dependent way. We name this attention mechanism Forgetting Attention and the resulting model the Forgetting Transformer (FoX). We show that FoX outperforms the Transformer on long-context language modeling, length extrapolation, and short-context downstream tasks, while performing on par with the Transformer on long-context downstream tasks. Moreover, it is compatible with the FlashAttention algorithm and does not require any positional embeddings. Several analyses, including the needle-in-the-haystack test, show that FoX also retains the Transformer's superior long-context capabilities over recurrent sequence models such as Mamba-2, HGRN2, and DeltaNet. We also introduce a "Pro" block design that incorporates some common architectural components in recurrent sequence models and find it significantly improves the performance of both FoX and the Transformer.Our code is available at https://github.com/zhixuan-lin/forgetting-transformer.

445. SAM-CP: Marrying SAM with Composable Prompts for Versatile Segmentation

链接：https://iclr.cc/virtual/2025/poster/29459 abstract： The Segment Anything model (SAM) has shown a generalized ability to group image pixels into patches, but applying it to semantic-aware segmentation still faces major challenges. This paper presents SAM-CP, a simple approach that establishes two types of composable prompts beyond SAM and composes them for

versatile segmentation. Specifically, given a set of classes (in texts) and a set of SAM patches, the Type-I prompt judges whether a SAM patch aligns with a text label, and the Type-II prompt judges whether two SAM patches with the same text label also belong to the same instance. To decrease the complexity in dealing with a large number of semantic classes and patches, we establish a unified framework that calculates the affinity between (semantic and instance) queries and SAM patches, and then merges patches with high affinity to the query. Experiments show that SAM-CP achieves semantic, instance, and panoptic segmentation in both open and closed domains. In particular, it achieves state-of-the-art performance in open-vocabulary segmentation. Our research offers a novel and generalized methodology for equipping vision foundation models like SAM with multi-grained semantic perception abilities. Codes are released on https://github.com/ucas-vg/SAM-CP.

446. Geometry Image Diffusion: Fast and Data-Efficient Text-to-3D with Image-Based Surface Representation

链接：https://iclr.cc/virtual/2025/poster/30259 abstract： Generating high-quality 3D objects from textual descriptions remains a challenging problem due to high computational costs, the scarcity of 3D data, and the complexity of 3D representations. We introduce Geometry Image Diffusion (GIMDiffusion), a novel Text-to-3D model that utilizes geometry images to efficiently represent 3D shapes using 2D images, thereby avoiding the need for complex 3D-aware architectures. By integrating a Collaborative Control mechanism, we exploit the rich 2D priors of existing Text-to-Image models, such as Stable Diffusion, to achieve strong generalization despite limited 3D training data. This allows us to use only high-quality training data while retaining compatibility with guidance techniques such as IPAdapter. GIMDiffusion enables the generation of 3D assets at speeds comparable to current Text-to-Image models, without being restricted to manifold meshes during either training or inference. We simultaneously generate a UV unwrapping for the objects, consisting of semantically meaningful parts as well as internal structures, enhancing both usability and versatility.

447. Test-time Alignment of Diffusion Models without Reward Over-optimization

链接：https://iclr.cc/virtual/2025/poster/27896 abstract： Diffusion models excel in generative tasks, but aligning them with specific objectives while maintaining their versatility remains challenging. Existing fine-tuning methods often suffer from reward over-optimization, while approximate guidance approaches fail to optimize target rewards effectively. Addressing these limitations, we propose a training-free, test-time method based on Sequential Monte Carlo (SMC) to sample from the reward-aligned target distribution. Our approach, tailored for diffusion sampling and incorporating tempering techniques, achieves comparable or superior target rewards to fine-tuning methods while preserving diversity and cross-reward generalization. We demonstrate its effectiveness in single-reward optimization, multi-objective scenarios, and online black-box optimization. This work offers a robust solution for aligning diffusion models with diverse downstream objectives without compromising their general capabilities. Code is available at https://github.com/krafton-ai/DAS.

448. Lightning-Fast Image Inversion and Editing for Text-to-Image Diffusion Models

链接：https://iclr.cc/virtual/2025/poster/28072 abstract： Diffusion inversion is the problem of taking an image and a text prompt that describes it and finding a noise latent that would generate the exact same image. Most current deterministic inversion techniques operate by approximately solving an implicit equation and may converge slowly or yield poor reconstructed images. We formulate the problem by finding the roots of an implicit equation and devlop a method to solve it efficiently. Our solution is based on Newton-Raphson (NR), a well-known technique in numerical analysis. We show that a vanilla application of NR is computationally infeasible while naively transforming it to a computationally tractable alternative tends to converge to out-of-distribution solutions, resulting in poor reconstruction and editing. We therefore derive an efficient guided formulation that fastly converges and provides high-quality reconstructions and editing. We showcase our method on real image editing with three popular open-sourced diffusion models: Stable Diffusion, SDXL-Turbo, and Flux with different deterministic schedulers. Our solution, Guided Newton-Raphson Inversion, inverts an image within 0.4 sec (on an A100 GPU) for few-step models (SDXL-Turbo and Flux.1),opening the door for interactive image editing. We further show improved results in image interpolation and generation of rare objects.

449. Adaptive Retention & Correction: Test-Time Training for Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30680 abstract： Continual learning, also known as lifelong learning or incremental learning, refers to the process by which a model learns from a stream of incoming data over time. A common problem in continual learning is the classification layer’s bias towards the most recent task. Traditionally, methods have relied on incorporating data from past tasks during training to mitigate this issue. However, the recent shift in continual learning to memory-free environments has rendered these approaches infeasible. In this study, we propose a solution focused on the testing phase. We first introduce a simple Out-of-Task Detection method, OTD, designed to accurately identify samples from past tasks during testing. Leveraging OTD, we then propose: (1) an Adaptive Retention mechanism for dynamically tuning the classifier layer on past task data; (2) an Adaptive Correction mechanism for revising predictions when the model classifies data from previous tasks into classes from the current task. We name our approach Adaptive Retention & Correction (ARC). While

designed for memory-free environments, ARC also proves effective in memorybased settings. Extensive experiments show that our proposed method can be plugged in to virtually any existing continual learning approach without requiring any modifications to its training procedure. Specifically, when integrated with state-of-the-art approaches, ARC achieves an average performance increase of 2.7% and 2.6% on the CIFAR-100 and Imagenet-R datasets, respectively

450. Efficient Model-Based Reinforcement Learning Through Optimistic Thompson Sampling

链接：https://iclr.cc/virtual/2025/poster/30164 abstract： Learning complex robot behavior through interactions with the environment necessitates principled exploration. Effective strategies should prioritize exploring regions of the state-action space that maximize rewards, with optimistic exploration emerging as a promising direction aligned with this idea and enabling sample-efficient reinforcement learning. However, existing methods overlook a crucial aspect: the need for optimism to be informed by a belief connecting the reward and state. To address this, we propose a practical, theoretically grounded approach to optimistic exploration based on Thompson sampling. Our approach is the first that allows for reasoning about joint uncertainty over transitions and rewards for optimistic exploration. We apply our method on a set of MuJoCo and VMAS continuous control tasks. Our experiments demonstrate that optimistic exploration significantly accelerates learning in environments with sparse rewards, action penalties, and difficult-to-explore regions. Furthermore, we provide insights into when optimism is beneficial and emphasize the critical role of model uncertainty in guiding exploration.

451. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability

链接：https://iclr.cc/virtual/2025/poster/27644 abstract： Retrieval-Augmented Generation (RAG) models are designed to incorporate external knowledge, reducing hallucinations caused by insufficient parametric (internal) knowledge. However, even with accurate and relevant retrieved content, RAG models can still produce hallucinations by generating outputs that conflict with the retrieved information. Detecting such hallucinations requires disentangling how Large Language Models (LLMs) balance external and parametric knowledge. Current detection methods often focus on one of these mechanisms or without decoupling their intertwined effects, making accurate detection difficult. In this paper, we investigate the internal mechanisms behind hallucinations in RAG scenarios. We discover hallucinations occur when the Knowledge FFNs in LLMs overemphasize parametric knowledge in the residual stream, while Copying Heads fail to effectively retain or integrate external knowledge from retrieved content. Based on these findings, we propose ReDeEP, a novel method that detects hallucinations by decoupling LLM’s utilization of external context and parametric knowledge. Our experiments show that ReDeEP significantly improves RAG hallucination detection accuracy. Additionally, we introduce AARF, which mitigates hallucinations by modulating the contributions of Knowledge FFNs and Copying Heads.

452. Rethinking Audio-Visual Adversarial Vulnerability from Temporal and Modality Perspectives

链接：https://iclr.cc/virtual/2025/poster/28927 abstract： While audio-visual learning equips models with a richer understanding of the real world by leveraging multiple sensory modalities, this integration also introduces new vulnerabilities to adversarial attacks. In this paper, we present a comprehensive study of the adversarial robustness of audio-visual models, considering both temporal and modality-specific vulnerabilities. We propose two powerful adversarial attacks: 1) a temporal invariance attack that exploits the inherent temporal redundancy across consecutive time segments and 2) a modality misalignment attack that introduces incongruence between the audio and visual modalities. These attacks are designed to thoroughly assess the robustness of audio-visual models against diverse threats. Furthermore, to defend against such attacks, we introduce a novel audio-visual adversarial training framework. This framework addresses key challenges in vanilla adversarial training by incorporating efficient adversarial perturbation crafting tailored to multi-modal data and an adversarial curriculum strategy. Extensive experiments in the Kinetics-Sounds dataset demonstrate that our proposed temporal and modality-based attacks in degrading model performance can achieve state-of-the-art performance, while our adversarial training defense largely improves the adversarial robustness as well as the adversarial training efficiency.

453. Preference Optimization for Reasoning with Pseudo Feedback

链接：https://iclr.cc/virtual/2025/poster/28622 abstract： Preference optimization techniques, such as Direct Preference Optimization (DPO), are frequently employed to enhance the reasoning capabilities of large language models (LLMs) in domains like mathematical reasoning and coding, typically following supervised fine-tuning. These methods rely on high-quality labels for reasoning tasks to generate preference pairs; however, the availability of reasoning datasets with human-verified labels is limited.In this study, we introduce a novel approach to generate pseudo feedback for reasoning tasks by framing the labeling of solutions to reason problems as an evaluation against associated \emph{test cases}. We explore two forms of pseudo feedback based on test cases: one generated by frontier LLMs and the other by extending self-consistency to multi-test-case.We conduct experiments on both mathematical reasoning and coding tasks using pseudo feedback for preference optimization, and observe improvements across both tasks. Specifically, using Mathstral-7B as our base model, we improve MATH results from 58.3 to 68.6, surpassing both NuminaMath-72B and GPT-4-Turbo-1106-preview. In GSM8K and College Math, our scores increase from 85.6 to 90.3 and from 34.3 to 42.3, respectively. Building on Deepseek-coder-7B-v1.5, we

achieve a score of 24.3 on LiveCodeBench (from 21.1), surpassing Claude-3-Haiku.

454. Theory on Mixture-of-Experts in Continual Learning

链接：https://iclr.cc/virtual/2025/poster/30816 abstract： Continual learning (CL) has garnered significant attention because of its ability to adapt to new tasks that arrive over time. Catastrophic forgetting (of old tasks) has been identified as a major issue in CL, as the model adapts to new tasks. The Mixture-of-Experts (MoE) model has recently been shown to effectively mitigate catastrophic forgetting in CL, by employing a gating network to sparsify and distribute diverse tasks among multiple experts. However, there is a lack of theoretical analysis of MoE and its impact on the learning performance in CL. This paper provides the first theoretical results to characterize the impact of MoE in CL via the lens of overparameterized linear regression tasks. We establish the benefit of MoE over a single expert by proving that the MoE model can diversify its experts to specialize in different tasks, while its router learns to select the right expert for each task and balance the loads across all experts. Our study further suggests an intriguing fact that the MoE in CL needs to terminate the update of the gating network after sufficient training rounds to attain system convergence, which is not needed in the existing MoE studies that do not consider the continual task arrival. Furthermore, we provide explicit expressions for the expected forgetting and overall generalization error to characterize the benefit of MoE in the learning performance in CL. Interestingly, adding more experts requires additional rounds before convergence, which may not enhance the learning performance. Finally, we conduct experiments on both synthetic and real datasets to extend these insights from linear models to deep neural networks (DNNs), which also shed light on the practical algorithm design for MoE in CL.

455. Self-Boosting Large Language Models with Synthetic Preference Data

链接：https://iclr.cc/virtual/2025/poster/30794 abstract： Through alignment with human preferences, Large Language Models (LLMs) have advanced significantly in generating honest, harmless, and helpful responses. However, collecting high-quality preference data is a resource-intensive and creativity-demanding process, especially for the continual improvement of LLMs. We introduce SynPO, a self-boosting paradigm that leverages synthetic preference data for model alignment. SynPO employs an iterative mechanism wherein a self-prompt generator creates diverse prompts, and a response improver refines model responses progressively. This approach trains LLMs to autonomously learn the generative rewards for their own outputs and eliminates the need for large-scale annotation of prompts and human preferences. After four SynPO iterations, Llama3-8B and Mistral-7B show significant enhancements in instruction-following abilities, achieving over 22.1% win rate improvements on AlpacaEval 2.0 and ArenaHard. Simultaneously, SynPO improves the general performance of LLMs on various tasks, validated by a 3.2 to 5.0 average score increase on the well-recognized Open LLM leaderboard.

456. STAFF: Speculative Coreset Selection for Task-Specific Fine-tuning

链接：https://iclr.cc/virtual/2025/poster/30360 abstract： Task-specific fine-tuning is essential for the deployment of large language models (LLMs), but it requires significant computational resources and time. Existing solutions have proposed coreset selection methods to improve data efficiency and reduce model training overhead, but they still have limitations: ❶ Overlooking valuable samples at high pruning rates, which degrades the coreset's performance.❷ Requiring high time overhead during coreset selection to fine-tune and evaluate the target LLM. In this paper, we introduce STAFF, a speculative coreset selection method. STAFF leverages a small model from the same family as the target LLM to efficiently estimate data scores and then verifies the scores on the target LLM to accurately identify and allocate more selection budget to important regions while maintaining coverage of easy regions. We evaluate STAFF on three LLMs and three downstream tasks and show that STAFF improves the performance of SOTA methods by up to 54.3% and reduces selection overhead by up to 70.5% at different pruning rates. Furthermore, we observe that the coreset selected by STAFF at low pruning rates (i.e., 20%) can even obtain better fine-tuning performance than the full dataset.

457. Adversarial Policy Optimization for Offline Preference-based Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30939 abstract： In this paper, we study offline preference-based reinforcement learning (PbRL), where learning is based on pre-collected preference feedback over pairs of trajectories. While offline PbRL has demonstrated remarkable empirical success, existing theoretical approaches face challenges in ensuring conservatism under uncertainty, requiring computationally intractable confidence set constructions. We address this limitation by proposing Adversarial Preference-based Policy Optimization (APPO), a computationally efficient algorithm for offline PbRL that guarantees sample complexity bounds without relying on explicit confidence sets. By framing PbRL as a two-player game between a policy and a model, our approach enforces conservatism in a tractable manner. Using standard assumptions on function approximation and bounded trajectory concentrability, we derive a sample complexity bound. To our knowledge, APPO is the first offline PbRL algorithm to offer both statistical efficiency and practical applicability. Experimental results on continuous control tasks demonstrate that APPO effectively learns from complex datasets, showing comparable performance with existing state-of-the-art methods.

458. Dynamic Assortment Selection and Pricing with Censored Preference Feedback

链接：https://iclr.cc/virtual/2025/poster/30462 abstract： In this study, we investigate the problem of dynamic multi-product selection and pricing by introducing a novel framework based on a *censored multinomial logit* (C-MNL) choice model. In this model, sellers present a set of products with prices, and buyers filter out products priced above their valuation, purchasing at most one product from the remaining options based on their preferences. The goal is to maximize seller revenue by dynamically adjusting product offerings and prices, while learning both product valuations and buyer preferences through purchase feedback. To achieve this, we propose a Lower Confidence Bound (LCB) pricing strategy. By combining this pricing strategy with either an Upper Confidence Bound (UCB) or Thompson Sampling (TS) product selection approach, our algorithms achieve regret bounds of $\tilde{O}(d^{\frac{3}{2}}\sqrt{T/\kappa})$ and $\tilde{O}(d^{2}\sqrt{T/\kappa})$, respectively. Finally, we validate the performance of our methods through simulations, demonstrating their effectiveness.

459. IntersectionZoo: Eco-driving for Benchmarking Multi-Agent Contextual Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29288 abstract： Despite the popularity of multi-agent reinforcement learning (RL) in simulated and two-player applications, its success in messy real-world applications has been limited. A key challenge lies in its generalizability across problem variations, a common necessity for many real-world problems. Contextual reinforcement learning (CRL) formalizes learning policies that generalize across problem variations. However, the lack of standardized benchmarks for multi-agent CRL has hindered progress in the field. Such benchmarks are desired to be based on real-world applications to naturally capture the many open challenges of real-world problems that affect generalization. To bridge this gap, we propose IntersectionZoo, a comprehensive benchmark suite for multi-agent CRL through the real-world application of cooperative eco-driving in urban road networks. The task of cooperative eco-driving is to control a fleet of vehicles to reduce fleet-level vehicular emissions. By grounding IntersectionZoo in a real-world application, we naturally capture real-world problem characteristics, such as partial observability and multiple competing objectives. IntersectionZoo is built on data-informed simulations of 16,334 signalized intersections derived from 10 major US cities, modeled in an open-source industry-grade microscopic traffic simulator. By modeling factors affecting vehicular exhaust emissions (e.g., temperature, road conditions, travel demand), IntersectionZoo provides one million data-driven traffic scenarios. Using these traffic scenarios, we benchmark popular multi-agent RL and human-like driving algorithms and demonstrate that the popular multi-agent RL algorithms struggle to generalize in CRL settings.

460. High-dimension Prototype is a Better Incremental Object Detection Learner

链接：https://iclr.cc/virtual/2025/poster/30884 abstract： Incremental object detection (IOD), surpassing simple classification, requires the simultaneous overcoming of catastrophic forgetting in both recognition and localization tasks, primarily due to the significantly higher feature space complexity. Integrating Knowledge Distillation (KD) would mitigate the occurrence of catastrophic forgetting. However, the challenge of knowledge shift caused by invisible previous task data hampers existing KD-based methods, leading to limited improvements in IOD performance. This paper aims to alleviate knowledge shift by enhancing the accuracy and granularity in describing complex high-dimensional feature spaces. To this end, we put forth a novel higher-dimension-prototype learning approach for KD-based IOD, enabling a more flexible, accurate, and fine-grained representation of feature distributions without the need to retain any previous task data. Existing prototype learning methods calculate feature centroids or statistical Gaussian distributions as prototypes, disregarding actual irregular distribution information or leading to inter-class feature overlap, which is not directly applicable to the more difficult task of IOD with complex feature space. To address the above issue, we propose a Gaussian Mixture Distribution-based Prototype (GMDP), which explicitly models the distribution relationships of different classes by directly measuring the likelihood of embedding from new and old models into class distribution prototypes in a higher dimension manner. Specifically, GMDP dynamically adapts the component weights and corresponding means/variances of class distribution prototypes to represent both intra-class and inter-class variability more accurately. Progressing into a new task, GMDP constrains the distance between the distribution of new and previous task classes, minimizing overlap with existing classes and thus striking a balance between stability and adaptability. GMDP can be readily integrated into existing IOD methods to enhance performance further. Extensive experiments on the PASCAL VOC and MS-COCO show that our method consistently exceeds four baselines by a large margin and significantly outperforms other SOTA results under various settings.

461. Diverse Preference Learning for Capabilities and Alignment

链接：https://iclr.cc/virtual/2025/poster/28303 abstract： As LLMs increasingly impact society, their ability to represent diverse perspectives is critical. However, recent studies reveal that alignment algorithms such as RLHF and DPO significantly reduce the diversity of LLM outputs. Not only do aligned LLMs generate text with repetitive structure and word choice, they also approach problems in more uniform ways, and their responses reflect a narrower range of societal perspectives. We attribute this problem to the KL divergence regularizer employed in preference learning algorithms. This causes the model to overweight majority opinions and sacrifice diversity in exchange for optimal reward. To address this, we propose Soft Preference Learning, which decouples the entropy and cross-entropy terms in the KL penalty — allowing for fine-grained control over LLM generation diversity. From a capabilities perspective, LLMs trained using Soft Preference Learning attain higher accuracy on difficult repeated sampling tasks and produce outputs with greater semantic and lexical diversity. From an alignment perspective, they are capable of representing a wider range of societal viewpoints and display improved logit calibration. Notably, Soft Preference Learning resembles, but is a Pareto improvement over, standard temperature scaling.

462. Topological Blindspots: Understanding and Extending Topological Deep Learning Through the Lens of Expressivity

链接：https://iclr.cc/virtual/2025/poster/30372 abstract： Topological deep learning (TDL) is a rapidly growing field that seeks to leverage topological structure in data and facilitate learning from data supported on topological objects, ranging from molecules to 3D shapes. Most TDL architectures can be unified under the framework of higher-order message-passing (HOMP), which generalizes graph message-passing to higher-order domains. In the first part of the paper, we explore HOMP's expressive power from a topological perspective, demonstrating the framework's inability to capture fundamental topological and metric invariants such as diameter, orientability, planarity, and homology. In addition, we demonstrate HOMP's limitations in fully leveraging lifting and pooling methods on graphs. To the best of our knowledge, this is the first work to study the expressivity of TDL from a topological perspective. In the second part of the paper, we develop two new classes of architectures -- multi-cellular networks (MCN) and scalable MCN (SMCN) -- which draw inspiration from expressive GNNs. MCN can reach full expressivity, but scaling it to large data objects can be computationally expansive. Designed as a more scalable alternative, SMCN still mitigates many of HOMP's expressivity limitations. Finally, we design new benchmarks for evaluating models based on their ability to learn topological properties of complexes. We then evaluate SMCN on these benchmarks as well as on real-world graph datasets, demonstrating improvements over both HOMP baselines and expressive graph methods, highlighting the value of expressively leveraging topological information.

463. Language Representations Can be What Recommenders Need: Findings and Potentials

链接：https://iclr.cc/virtual/2025/poster/28934 abstract： Recent studies empirically indicate that language models (LMs) encode rich world knowledge beyond mere semantics, attracting significant attention across various fields.However, in the recommendation domain, it remains uncertain whether LMs implicitly encode user preference information. Contrary to prevailing understanding that LMs and traditional recommenders learn two distinct representation spaces due to the huge gap in language and behavior modeling objectives, this work re-examines such understanding and explores extracting a recommendation space directly from the language representation space.Surprisingly, our findings demonstrate that item representations, when linearly mapped from advanced LM representations, yield superior recommendation performance.This outcome suggests the possible homomorphism between the advanced language representation space and an effective item representation space for recommendation, implying that collaborative signals may be implicitly encoded within LMs.Motivated by the finding of homomorphism, we explore the possibility of designing advanced collaborative filtering (CF) models purely based on language representations without ID-based embeddings.To be specific, we incorporate several crucial components (i.e., a multilayer perceptron (MLP), graph convolution, and contrastive learning (CL) loss function) to build a simple yet effective model, with the language representations of item textual metadata (i.e., title) as the input.Empirical results show that such a simple model can outperform leading ID-based CF models on multiple datasets, which sheds light on using language representations for better recommendation.Moreover, we systematically analyze this simple model and find several key features for using advanced language representations:a good initialization for item representations, superior zero-shot recommendation abilities in new datasets, and being aware of user intention.Our findings highlight the connection between language modeling and behavior modeling, which can inspire both natural language processing and recommender system communities.

464. Causal Effect Estimation with Mixed Latent Confounders and Post-treatment Variables

链接：https://iclr.cc/virtual/2025/poster/28242 abstract： Causal inference from observational data has attracted considerable attention among researchers. One main obstacle is the handling of confounders. As direct measurement of confounders may not be feasible, recent methods seek to address the confounding bias via proxy variables, i.e., covariates postulated to be conducive to the inference of latent confounders. However, the selected proxies may scramble both confounders and post-treatment variables in practice, which risks biasing the estimation by controlling for variables affected by the treatment. In this paper, we systematically investigate the bias due to latent post-treatment variables, i.e., latent post-treatment bias, in causal effect estimation. Specifically, we first derive the bias when selected proxies scramble both latent confounders and post-treatment variables, which we demonstrate can be arbitrarily bad. We then propose a Confounder-identifiable VAE (CiVAE) to address the bias. Based on a mild assumption that the prior of latent variables that generate the proxy belongs to a general exponential family with at least one invertible sufficient statistic in the factorized part, CiVAE individually identifies latent confounders and latent post-treatment variables up to bijective transformations. We then prove that with individual identification, the intractable disentanglement problem of latent confounders and post-treatment variables can be transformed into a tractable independence test problem despite arbitrary dependence may exist among them. Finally, we prove that the true causal effects can be unbiasedly estimated with transformed confounders inferred by CiVAE. Experiments on both simulated and real-world datasets demonstrate significantly improved robustness of CiVAE.

465. Fast and Accurate Blind Flexible Docking

链接：https://iclr.cc/virtual/2025/poster/28683 abstract： Molecular docking that predicts the bound structures of small molecules (ligands) to their protein targets, plays a vital role in drug discovery. However, existing docking methods often face limitations: they either overlook crucial structural changes by assuming protein rigidity or suffer from low computational efficiency

due to their reliance on generative models for structure sampling. To address these challenges, we propose FABFlex, a fast and accurate regression-based multi-task learning model designed for realistic blind flexible docking scenarios, where proteins exhibit flexibility and binding pocket sites are unknown (blind). Specifically, FABFlex's architecture comprises three specialized modules working in concert: (1) A pocket prediction module that identifies potential binding sites, addressing the challenges inherent in blind docking scenarios. (2) A ligand docking module that predicts the bound (holo) structures of ligands from their unbound (apo) states. (3) A pocket docking module that forecasts the holo structures of protein pockets from their apo conformations. Notably, FABFlex incorporates an iterative update mechanism that serves as a conduit between the ligand and pocket docking modules, enabling continuous structural refinements. This approach effectively integrates the three subtasks of blind flexible docking—pocket identification, ligand conformation prediction, and protein flexibility modeling—into a unified, coherent framework. Extensive experiments on public benchmark datasets demonstrate that FABFlex not only achieves superior effectiveness in predicting accurate binding modes but also exhibits a significant speed advantage (208$\times$) compared to existing state-of-the-art methods. Our code is released at~\url{https://github.com/tmlr-group/FABFlex}.

466. 3D-MolT5: Leveraging Discrete Structural Information for Molecule-Text Modeling

链接：https://iclr.cc/virtual/2025/poster/28938 abstract： The integration of molecular and natural language representations has emerged as a focal point in molecular science, with recent advancements in Language Models (LMs) demonstrating significant potential for comprehensive modeling of both domains. However, existing approaches face notable limitations, particularly in their neglect of three-dimensional (3D) information, which is crucial for understanding molecular structures and functions. While some efforts have been made to incorporate 3D molecular information into LMs using external structure encoding modules, significant difficulties remain, such as insufficient interaction across modalities in pre-training and challenges in modality alignment. To address the limitations, we propose \textbf{3D-MolT5}, a unified framework designed to model molecule in both sequence and 3D structure spaces. The key innovation of our approach lies in mapping fine-grained 3D substructure representations into a specialized 3D token vocabulary. This methodology facilitates the seamless integration of sequence and structure representations in a tokenized format, enabling 3D-MolT5 to encode molecular sequences, molecular structures, and text sequences within a unified architecture. Leveraging this tokenized input strategy, we build a foundation model that unifies the sequence and structure data formats. We then conduct joint pre-training with multi-task objectives to enhance the model's comprehension of these diverse modalities within a shared representation space. Thus, our approach significantly improves cross-modal interaction and alignment, addressing key challenges in previous work. Further instruction tuning demonstrated that our 3D-MolT5 has strong generalization ability and surpasses existing methods with superior performance in multiple downstream tasks, such as nearly 70\% improvement on the molecular property prediction task compared to state-of-the-art methods. Our code is available at \url{https://github.com/QizhiPei/3D-MolT5}.

467. A Simple yet Effective $\Delta\Delta G$ Predictor is An Unsupervised Antibody Optimizer and Explainer

链接：https://iclr.cc/virtual/2025/poster/30140 abstract： The proteins that exist today have been optimized over billions of years of natural evolution, during which nature creates random mutations and selects them. The discovery of functionally promising mutations is challenged by the limited evolutionary accessible regions, i.e., only a small region on the fitness landscape is beneficial. There have been numerous priors used to constrain protein evolution to regions of landscapes with high-fitness variants, among which the change in binding free energy ($\Delta\Delta G$) of protein complexes upon mutations is one of the most commonly used priors. However, the huge mutation space poses two challenges: (1) how to improve the efficiency of $\Delta\Delta G$ prediction for fast mutation screening; and (2) how to explain mutation preferences and efficiently explore accessible evolutionary regions. To address these challenges, we propose a lightweight $\Delta\Delta G$ predictor (Light-DDG), which adopts a structure-aware Transformer as the backbone and enhances it by knowledge distilled from existing powerful but computationally heavy $\Delta\Delta G$ predictors. Additionally, we augmented, annotated, and released a large-scale dataset containing millions of mutation data for pre-training Light-DDG. We find that such a simple yet effective Light-DDG can serve as a good unsupervised antibody optimizer and explainer. For the target antibody, we propose a novel Mutation Explainer to learn mutation preferences, which accounts for the marginal benefit of each mutation per residue. To further explore accessible evolutionary regions, we conduct preference-guided antibody optimization and evaluate antibody candidates quickly using Light-DDG to identify desirable mutations. Extensive experiments have demonstrated the effectiveness of Light-DDG in terms of test generalizability, noise robustness, and inference practicality, e.g., 89.7$\times$ inference acceleration and 15.45\% performance gains over previous state-of-the-art baselines. A case study of SARS-CoV-2 further demonstrates the crucial role of Light-DDG for mutation explanation and antibody optimization.

468. Efficient Source-Free Time-Series Adaptation via Parameter Subspace Disentanglement

链接：https://iclr.cc/virtual/2025/poster/29710 abstract： In this paper, we propose a framework for efficient Source-Free Domain Adaptation (SFDA) in the context of time-series, focusing on enhancing both parameter efficiency and data-sample utilization. Our approach introduces an improved paradigm for source-model preparation and target-side adaptation, aiming to enhance training efficiency during target adaptation. Specifically, we reparameterize the source model's weights in a Tucker-style decomposed manner, factorizing the model into a compact form during the source model preparation phase. During target-side adaptation, only a subset of these decomposed factors is fine-tuned, leading to significant improvements in training

efficiency. We demonstrate using PAC Bayesian analysis that this selective fine-tuning strategy implicitly regularizes the adaptation process by constraining the model's learning capacity. Furthermore, this re-parameterization reduces the overall model size and enhances inference efficiency, making the approach particularly well suited for resource-constrained devices. Additionally, we demonstrate that our framework is compatible with various SFDA methods and achieves significant computational efficiency, reducing the number of fine-tuned parameters and inference overhead in terms of MACs by over 90\% while maintaining model performance.

469. Reconstructive Visual Instruction Tuning

链接：https://iclr.cc/virtual/2025/poster/30741 abstract： This paper introduces reconstructive visual instruction tuning (ROSS), a family of Large Multimodal Models (LMMs) that exploit vision-centric supervision signals. In contrast to conventional visual instruction tuning approaches that exclusively supervise text outputs, ROSS prompts LMMs to supervise visual outputs via reconstructing input images. By doing so, it capitalizes on the inherent richness and detail present within input images themselves, which are often lost in pure text supervision. However, producing meaningful feedback from natural images is challenging due to the heavy spatial redundancy of visual signals. To address this issue, ROSS employs a denoising objective to reconstruct latent representations of input images, avoiding directly regressing exact raw RGB values. This intrinsic activation design inherently encourages LMMs to maintain image detail, thereby enhancing their fine-grained comprehension capabilities and reducing hallucinations. Empirically, ROSS consistently brings significant improvements across different visual encoders and language models. In comparison with extrinsic assistance state-of-the-art alternatives that aggregate multiple visual experts, ROSS delivers competitive performance with a single SigLIP visual encoder, demonstrating the efficacy of our vision-centric supervision tailored for visual outputs. The code will be made publicly available upon acceptance.

470. Cafe-Talk: Generating 3D Talking Face Animation with Multimodal Coarse- and Fine-grained Control

链接：https://iclr.cc/virtual/2025/poster/29627 abstract： Speech-driven 3D talking face method should offer both accurate lip synchronization and controllable expressions. Previous methods solely adopt discrete emotion labels to globally control expressions throughout sequences while limiting flexible fine-grained facial control within the spatiotemporal domain. We propose a diffusion-transformer-based 3D talking face generation model, Cafe-Talk, which simultaneously incorporates coarse- and fine-grained multimodal control conditions. Nevertheless, the entanglement of multiple conditions challenges achieving satisfying performance. To disentangle speech audio and fine-grained conditions, we employ a two-stage training pipeline. Specifically, Cafe-Talk is initially trained using only speech audio and coarse-grained conditions. Then, a proposed fine-grained control adapter gradually adds fine-grained instructions represented by action units (AUs), preventing unfavorable speech-lip synchronization. To disentangle coarse- and fine-grained conditions, we design a swap-label training mechanism, which enables the dominance of the fine-grained conditions. We also devise a mask-based CFG technique to regulate the occurrence and intensity of fine-grained control. In addition, a text-based detector is introduced with text-AU alignment to enable natural language user input and further support multimodal control. Extensive experimental results prove that Cafe-Talk achieves state-of-the-art lip synchronization and expressiveness performance and receives wide acceptance in fine-grained control in user studies.

471. Sequential Stochastic Combinatorial Optimization Using Hierarchal Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30618 abstract： Reinforcement learning (RL) has emerged as a promising tool for combinatorial optimization (CO) problems due to its ability to learn fast, effective, and generalizable solutions. Nonetheless, existing works mostly focus on one-shot deterministic CO, while sequential stochastic CO (SSCO) has rarely been studied despite its broad applications such as adaptive influence maximization (IM) and infectious disease intervention. In this paper, we study the SSCO problem where we first decide the budget (e.g., number of seed nodes in adaptive IM) allocation for all time steps, and then select a set of nodes for each time step. The few existing studies on SSCO simplify the problems by assuming a uniformly distributed budget allocation over the time horizon, yielding suboptimal solutions. We propose a generic hierarchical RL (HRL) framework called wake-sleep option (WS-option), a two-layer option-based framework that simultaneously decides adaptive budget allocation on the higher layer and node selection on the lower layer. WS-option starts with a coherent formulation of the two-layer Markov decision processes (MDPs), capturing the interdependencies between the two layers of decisions. Building on this, WS-option employs several innovative designs to balance the model's training stability and computational efficiency, preventing the vicious cyclic interference issue between the two layers. Empirical results show that WS-option exhibits significantly improved effectiveness and generalizability compared to traditional methods. Moreover, the learned model can be generalized to larger graphs, which significantly reduces the overhead of computational resources.

472. Optimizing Neural Network Representations of Boolean Networks

链接：https://iclr.cc/virtual/2025/poster/31213 abstract： Neural networks are known to be universal computers for Boolean functions. Recent advancements in hardware have significantly reduced matrix multiplication times, making neural network simulation both fast and efficient. Consequently, functions defined by complex Boolean networks are increasingly viable candidates for simulation through their neural network representation. Prior research has introduced a general method for deriving neural network representations of Boolean networks. However, the resulting neural networks are often suboptimal in

terms of the number of neurons and connections, leading to slower simulation performance. Optimizing them while preserving functional equivalence --lossless optimization-- is an NP-hard problem, and current methods only provide lossy solutions. In this paper, we present a deterministic algorithm to optimize such neural networks in terms of neurons and connections while preserving functional equivalence. Moreover, to accelerate the compression of the neural network, we introduce an objective-aware algorithm that exploits representations that are shared among subproblems of the overall optimization. We demonstrate experimentally that we are able to reduce connections and neurons by up to 70% and 60%, respectively, in comparison to state-of-the-art. We also find that our objective-aware algorithm results in consistent speedups in optimization time, achieving up to 34.3x and 5.9x speedup relative to naive and caching solutions, respectively. Our methods are of practical relevance to applications such as high-throughput circuit simulation and placing neurosymbolic systems on the same hardware architecture.

473. ExACT: Teaching AI Agents to Explore with Reflective-MCTS and Exploratory Learning

链接：https://iclr.cc/virtual/2025/poster/30295 abstract： Autonomous agents have demonstrated significant potential in automating complex multistep decision-making tasks. However, even state-of-the-art vision-language models (VLMs), such as GPT-4o, still fall short of human-level performance, particularly in intricate web environments and long-horizon planning tasks. To address these limitations, we introduce Reflective Monte Carlo Tree Search (R-MCTS), a novel test-time algorithm designed to enhance the ability of AI agents, e.g., powered by GPT-4o, to explore decision space on the fly.R-MCTS extends traditional MCTS by 1) incorporating contrastive reflection, allowing agents to learn from past interactions and dynamically improve their search efficiency; and 2) using multi-agent debate to provide reliable state evaluation. Moreover, we improve the agent's performance by fine-tuning GPT-4o through self-learning, using R-MCTS generated tree traversals without any human-provided labels. On the challenging VisualWebArena benchmark, our GPT-4o-based R-MCTS agent achieves a 6% to 30% relative improvement across various tasks compared to the previous state-of-the-art. Additionally, we show that the knowledge gained from test-time search can be effectively transferred back to GPT-4o via fine-tuning. The fine-tuned GPT-4o matches 97\% of R-MCTS's performance while reducing compute usage by a factor of four at test time. Furthermore, qualitative results reveal that the fine-tuned GPT-4o model demonstrates the ability to explore the environment, evaluate a state, and backtrack to viable ones when it detects that the current state cannot lead to success. Moreover, our work demonstrates the compute scaling properties in both training - data collection with R-MCTS - and testing time. These results suggest a promising research direction to enhance VLMs' reasoning and planning capabilities for agentic applications via test-time search and self-learning.

474. CBGBench: Fill in the Blank of Protein-Molecule Complex Binding Graph

链接：https://iclr.cc/virtual/2025/poster/28468 abstract： Structure-based drug design (SBDD) aims to generate potential drugs that can bind to a target protein and is greatly expedited by the aid of AI techniques in generative models. However, a lack of systematic understanding persists due to the diverse settings, complex implementation, difficult reproducibility, and task singularity. Firstly, the absence of standardization can lead to unfair comparisons and inconclusive insights. To address this dilemma, we propose CBGBench, a comprehensive benchmark for SBDD, that unifies the task as a generative graph completion, analogous to fill-in-the-blank of the 3D complex binding graph. By categorizing existing methods based on their attributes, CBGBench facilitates a modular and extensible framework that implements cutting-edge methods. Secondly, a single de novo molecule generation task can hardly reflect their capabilities. To broaden the scope, we adapt these models to a range of tasks essential in drug design, considered sub-tasks within the graph fill-in-the-blank tasks. These tasks include the generative designation of de novo molecules, linkers, fragments, scaffolds, and sidechains, all conditioned on the structures of protein pockets. Our evaluations are conducted with fairness, encompassing comprehensive perspectives on interaction, chemical properties, geometry authenticity, and substructure validity. We further provide insights with analysis from empirical studies. Our results indicate that there is potential for further improvements on many tasks, with optimization in network architectures, and effective incorporation of chemical prior knowledge. Finally, to lower the barrier to entry and facilitate further developments in the field, we also provide a single codebase that unifies the discussed models, data pre-processing, training, sampling, and evaluation.

475. Block Diffusion: Interpolating Between Autoregressive and Diffusion Language Models

链接：https://iclr.cc/virtual/2025/poster/28004 abstract： Diffusion language models offer unique benefits over autoregressive models due to their potential for parallelized generation and controllability, yet they lag in likelihood modeling and are limited to fixed-length generation. In this work, we introduce a class of block diffusion language models that interpolate between discrete denoising diffusion and autoregressive models. Block diffusion overcomes key limitations of both approaches by supporting flexible-length generation and improving inference efficiency with KV caching and parallel token sampling. We propose a recipe for building effective block diffusion models that includes an efficient training algorithm, estimators of gradient variance, and data-driven noise schedules to minimize the variance. Block diffusion sets a new state-of-the-art performance among diffusion models on language modeling benchmarks and enables generation of arbitrary-length sequences. We provide the code, along with the model weights and blog post on the project page: https://m-arriola.com/bd3lms/

476. R^2-Guard: Robust Reasoning Enabled LLM Guardrail via

Knowledge-Enhanced Logical Reasoning

链接：https://iclr.cc/virtual/2025/poster/30497 abstract： As large language models (LLMs) become increasingly prevalent across various applications, it is critical to establish safety guardrails to moderate input/output content of LLMs and ensure compliance with safety policies. Existing guardrail models, such as OpenAI Mod and LlamaGuard, treat various safety categories (e.g., self-harm, self-harm/instructions) independently and fail to explicitly capture the intercorrelations among them. This has led to limitations such as ineffectiveness due to inadequate training on long-tail data from correlated safety categories, susceptibility to jailbreaking attacks, and inflexibility regarding new safety categories.To address these limitations, we propose R^2-Guard, a robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning. Specifically, R^2-Guard comprises two parts: data-driven guardrail models and reasoning components. The data-driven guardrail models provide unsafety probabilities of moderated content on different safety categories.We then encode safety knowledge among different categories as first-order logical rules and embed them into a probabilistic graphic model (PGM) based reasoning component. The unsafety probabilities of different categories from data-driven guardrail models are sent to the reasoning component for final inference. We employ two types of PGMs: Markov logic networks (MLNs) and probabilistic circuits (PCs), and optimize PCs to achieve precision-efficiency balance via improved graph structure. We also propose different methods to optimize the weights of knowledge. To further perform stress tests for guardrail models, we employ a pairwise construction method to construct a new safety benchmark TwinSafety, which features principled categories and presents new challenges for moderation. We show that R^2-Guard is effective even given unrepresentative categories or challenging jailbreaking prompts. We demonstrate the effectiveness of R^2-Guard by comparisons with eight strong guardrail models on six standard moderation datasets, and demonstrate the robustness of R^2-Guard against four SOTA jailbreaking attacks. R^2-Guard significantly surpasses SOTA method LlamaGuard by 12.6% on standard moderation datasets and by 59.9% against jailbreaking attacks.We further reveal that R^2-Guard can effectively adapt to safety category updates by simply editing the PGM reasoning graph.

477. AdvWave: Stealthy Adversarial Jailbreak Attack against Large Audio-Language Models

链接：https://iclr.cc/virtual/2025/poster/31267 abstract： Recent advancements in large audio-language models (LALMs) have enabled speech-based user interactions, significantly enhancing user experience and accelerating the deployment of LALMs in real-world applications. However, ensuring the safety of LALMs is crucial to prevent risky outputs that may raise societal concerns or violate AI regulations. Despite the importance of this issue, research on jailbreaking LALMs remains limited due to their recent emergence and the additional technical challenges they present compared to attacks on DNN-based audio models. Specifically, the audio encoders in LALMs, which involve discretization operations, often lead to gradient shattering, hindering the effectiveness of attacks relying on gradient-based optimizations. The behavioral variability of LALMs further complicates the identification of effective (adversarial) optimization targets. Moreover, enforcing stealthiness constraints on adversarial audio waveforms introduces a reduced, non-convex feasible solution space, further intensifying the challenges of the optimization process. To overcome these challenges, we develop AdvWave, the first jailbreak framework against LALMs. We propose a dual-phase optimization method that addresses gradient shattering, enabling effective end-to-end gradient-based optimization. Additionally, we develop an adaptive adversarial target search algorithm that dynamically adjusts the adversarial optimization target based on the response patterns of LALMs for specific queries. To ensure that adversarial audio remains perceptually natural to human listeners, we design a classifier-guided optimization approach that generates adversarial noise resembling common urban sounds. Extensive evaluations on multiple advanced LALMs demonstrate that AdvWave outperforms baseline methods, achieving a 40\% higher average jailbreak attack success rate. Both audio stealthiness metrics and human evaluations confirm that adversarial audio generated by AdvWave is indistinguishable from natural sounds. We believe AdvWave will inspire future research aiming to enhance the safety alignment of LALMs, supporting their responsible deployment in real-world scenarios.

478. Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse

链接：https://iclr.cc/virtual/2025/poster/30139 abstract： LLMs are an integral component of retrieval-augmented generation (RAG) systems. While many studies focus on evaluating the overall quality of end-to-end RAG systems, there is a gap in understanding the appropriateness of LLMs for the RAG task. To address this, we introduce Trust-Score, a holistic metric that evaluates the trustworthiness of LLMs within the RAG framework. Our results show that various prompting methods, such as in-context learning, fail to effectively adapt LLMs to the RAG task as measured by Trust-Score. Consequently, we propose Trust-Align, a method to align LLMs for improved Trust-Score performance. 26 out of 27 models aligned using Trust-Align substantially outperform competitive baselines on ASQA, QAMPARI, and ELI5. Specifically, in LLaMA-3-8b, Trust-Align outperforms FRONT on ASQA (↑12.56), QAMPARI (↑36.04), and ELI5 (↑17.69). Trust-Align also significantly enhances models' ability to correctly refuse and provide quality citations. We also demonstrate the effectiveness of Trust-Align across different open-weight models, including the LLaMA series (1b to 8b), Qwen-2.5 series (0.5b to 7b), and Phi3.5 (3.8b). We release our code at https://github.com/declare-lab/trust-align.

479. On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization

链接：https://iclr.cc/virtual/2025/poster/31499 abstract： Adaptive gradient methods are workhorses in deep learning. However, the convergence guarantees of adaptive gradient methods for nonconvex optimization have not been thoroughly studied. In this paper, we provide a fine-grained convergence analysis for a general class of adaptive gradient methods including AMSGrad, RMSProp and AdaGrad. For smooth nonconvex functions, we prove that adaptive gradient methods in expectation converge to a first-order stationary point. Our convergence rate is better than existing results for adaptive gradient methods in terms of dimension. In addition, we also prove high probability bounds on the convergence rates of AMSGrad, RMSProp as well as AdaGrad, which have not been established before. Our analyses shed light on better understanding the mechanism behind adaptive gradient methods in optimizing nonconvex objectives.

480. Time After Time: Deep-Q Effect Estimation for Interventions on When and What to do

链接：https://iclr.cc/virtual/2025/poster/30918 abstract： Problems in fields such as healthcare, robotics, and finance requires reasoning about the value both of what decision or action to take and when to take it. The prevailing hope is that artificial intelligence will support such decisions by estimating the causal effect of policies such as how to treat patients or how to allocate resources over time. However, existing methods for estimating the effect of a policy struggle with \emph{irregular time}. They either discretize time, or disregard the effect of timing policies.We present a new deep-Q algorithm that estimates the effect of both when and what to do called Earliest Disagreement Q-Evaluation (EDQ). EDQ makes use of recursion for the Q-function that is compatible with flexible sequence models, such as transformers. EDQ provides accurate estimates under standard assumptions. We validate the approach through experiments on survival time and tumor growth tasks.

481. Can Neural Networks Achieve Optimal Computational-statistical Tradeoff? An Analysis on Single-Index Model

链接：https://iclr.cc/virtual/2025/poster/28668 abstract： In this work, we tackle the following question: Can neural networks trained with gradient-based methods achieve the optimal statistical-computational tradeoff in learning Gaussian single-index models? Prior research has shown that any polynomial-time algorithm under the statistical query (SQ) framework requires $\Omega(d^{s^\star/2}\lor d)$ samples, where $s^\star$ is the generative exponent representing the intrinsic difficulty of learning the underlying model.However, it remains unknown whether neural networks can achieve this sample complexity. Inspired by prior techniques such as label transformation and landscape smoothing for learning single-index models, we propose a unified gradient-based algorithm for training a two-layer neural network in polynomial time.Our method is adaptable to a variety of loss and activation functions, covering a broad class of existing approaches.We show that our algorithm learns a feature representation that strongly aligns with the unknown signal $\theta^\star$, with sample complexity $\tilde O (d^{s^\star/2} \lor d)$, matching the SQ lower bound up to a polylogarithmic factor for all generative exponents $s^\star\geq 1$.Furthermore, we extend our approach to the setting where $\theta^\star$ is k-sparse for $k = o(\sqrt{d})$ by introducing a novel weight perturbation technique that leverages the sparsity structure. We derive a corresponding SQ lower bound of order $\tilde\Omega(k^{s^\star})$, matched by our method up to a polylogarithmic factor.Our framework, especially the weight perturbation technique, is of independent interest, and suggests potential gradient-based solutions to other problems such as sparse tensor PCA.

482. Learning Regularized Graphon Mean-Field Games with Unknown Graphons

链接：https://iclr.cc/virtual/2025/poster/31383 abstract： We design and analyze reinforcement learning algorithms for Graphon Mean-Field Games (GMFGs). In contrast to previous works that require the precise values of the graphons, we aim to learn the Nash Equilibrium (NE) of the regularized GMFGs when the graphons are unknown. Our contributions are threefold. First, we propose the Proximal Policy Optimization for GMFG (GMFG-PPO) algorithm and show that it converges at a rate of $\tilde{O}(T^{-1/3})$ after T iterations with an estimation oracle, improving on a previous work by Xie et al. (ICML, 2021). Second, using kernel embedding of distributions, we design efficient algorithms to estimate the transition kernels, reward functions, and graphons from sampled agents. Convergence rates are then derived when the positions of the agents are either known or unknown. Results for the combination of the optimization algorithm GMFG-PPO and the estimation algorithm are then provided. These algorithms are the first specifically designed for learning graphons from sampled agents. Finally, the efficacy of the proposed algorithms are corroborated through simulations. These simulations demonstrate that learning the unknown graphons reduces the exploitability effectively.

483. Q-Adapter: Customizing Pre-trained LLMs to New Preferences with Forgetting Mitigation

链接：https://iclr.cc/virtual/2025/poster/29368 abstract： Large Language Models (LLMs), trained on a large amount of corpus, have demonstrated remarkable abilities. However, it may not be sufficient to directly apply open-source LLMs like Llama to certain real-world scenarios, since most of them are trained for \emph{general} purposes. Thus, the demands for customizing publicly available LLMs emerge, but are currently under-studied. In this work, we consider customizing pre-trained LLMs with new human preferences. Specifically, the LLM should not only meet the new preference but also preserve its original capabilities after customization. Drawing inspiration from the observation that human preference can be expressed as a reward model, we propose to cast LLM customization as optimizing the sum of two reward functions, one of which (denoted as r_1) was used to

pre-train the LLM while the other (denoted as r_2) characterizes the new human preference. The obstacle here is that both reward functions are unknown, making the application of modern reinforcement learning methods infeasible. Thanks to the residual Q-learning framework, we can restore the customized LLM with the pre-trained LLM and the \emph{residual Q-function} without the reward function r_1. Moreover, we find that for a fixed pre-trained LLM, the reward function r_2 can be derived from the residual Q-function, enabling us to directly learn the residual Q-function from the new human preference data upon the Bradley-Terry model. We name our method Q-Adapter as it introduces an adapter module to approximate the residual Q-function for customizing the pre-trained LLM towards the new preference. Experiments based on the Llama-3.1 model on the DSP dataset and HH-RLHF dataset illustrate the superior effectiveness of Q-Adapter on both retaining existing knowledge and learning new preferences. Our code is available at \url{https://github.com/LAMDA-RL/Q-Adapter}.

484. Online Preference Alignment for Language Models via Count-based Exploration

链接：https://iclr.cc/virtual/2025/poster/29036 abstract： Reinforcement Learning from Human Feedback (RLHF) has shown great potential in fine-tuning Large Language Models (LLMs) to align with human preferences. Existing methods perform preference alignment from a fixed dataset, which can be limited in data coverage and the resulting reward model is hard to generalize in out-of-distribution responses. Thus, online RLHF is more desirable to empower the LLM to explore outside the support of the initial dataset by iteratively collecting the prompt-response pairs. In this paper, we study the fundamental problem in online RLHF, i.e., how to explore for LLM. We give a theoretical motivation in linear reward assumption to show that an optimistic reward with an upper confidence bound (UCB) term leads to a provably efficient RLHF policy. Then, we reformulate our objective to direct preference optimization with an exploration term, where the UCB-term can be converted to a count-based exploration bonus. We further propose a practical algorithm, named Count-based Online Preference Optimization (COPO), which leverages a simple coin-flip counting module to estimate the pseudo-count of a prompt-response pair in previously collected data. COPO encourages LLMs to balance exploration and preference optimization in an iterative manner, which enlarges the exploration space and the entire data coverage of iterative LLM policies. We conduct online RLHF experiments on Zephyr and Llama-3 models. The results on instruction-following and standard academic benchmarks show that COPO significantly increases performance.

485. Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization

链接：https://iclr.cc/virtual/2025/poster/27968 abstract： Direct Preference Optimization (DPO) and its variants are increasingly used for aligning language models with human preferences. Although these methods are designed to teach a model to generate preferred responses more frequently relative to dispreferred responses, prior work has observed that the likelihood of preferred responses often decreases during training. The current work sheds light on the causes and implications of this counterintuitive phenomenon, which we term *likelihood displacement*. We demonstrate that likelihood displacement can be *catastrophic*, shifting probability mass from preferred responses to responses with an opposite meaning. As a simple example, training a model to prefer $\texttt{No}$ over $\texttt{Never}$ can sharply increase the probability of $\texttt{Yes}$. Moreover, when aligning the model to refuse unsafe prompts, we show that such displacement can *unintentionally lead to unalignment*, by shifting probability mass from preferred refusal responses to harmful responses (e.g., reducing the refusal rate of Llama-3-8B-Instruct from 74.4% to 33.4%). We theoretically characterize that likelihood displacement is driven by preferences that induce similar embeddings, as measured by a *centered hidden embedding similarity (CHES)* score. Empirically, the CHES score enables identifying which training samples contribute most to likelihood displacement in a given dataset. Filtering out these samples effectively mitigated unintentional unalignment in our experiments. More broadly, our results highlight the importance of curating data with sufficiently distinct preferences, for which we believe the CHES score may prove valuable.

486. SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27805 abstract： Multimodal Large Language Models (MLLMs) are advancing the ability to reason about complex sports scenarios by integrating textual and visual information. To comprehensively evaluate their capabilities, we introduce SPORTU, a benchmark designed to assess MLLMs across multi-level sports reasoning tasks. SPORTU comprises two key components: SPORTU-text, featuring 900 multiple-choice questions with human-annotated explanations for rule comprehension and strategy understanding. This component focuses on testing models' ability to reason about sports solely through question-answering (QA), without requiring visual inputs; SPORTU-video, consisting of 1,701 slow-motion video clips across 7 different sports and 12,048 QA pairs, designed to assess multi-level reasoning, from simple sports recognition to complex tasks like foul detection and rule application. We evaluated four prevalent LLMs mainly utilizing few-shot learning paradigms supplemented by chain-of-thought (CoT) prompting on the SPORTU-text part. GPT-4o achieves the highest accuracy of 71\%, but still falls short of human-level performance, highlighting room for improvement in rule comprehension and reasoning. The evaluation for the SPORTU-video part includes 6 proprietary and 8 open-source MLLMs. Experiments show that models fall short on hard tasks that require deep reasoning and rule-based understanding. GPT-4o performs the best with only 57.8\% accuracy on the hard task, showing large room for improvement. We hope that SPORTU will serve as a critical step toward evaluating models' capabilities in sports understanding and reasoning. The dataset is available at https://github.com/chili-lab/SPORTU.

487. Can Video LLMs Refuse to Answer? Alignment for Answerability in Video Large Language Models

链接：https://iclr.cc/virtual/2025/poster/29769 abstract： In the broader context of deep learning, Multimodal Large Language Models have achieved significant breakthroughs by leveraging powerful Large Language Models as a backbone to align different modalities into the language space. A prime exemplification is the development of Video Large Language Models (Video-LLMs). While numerous advancements have been proposed to enhance the video understanding capabilities of these models, they are predominantly trained on questions generated directly from video content. However, in real-world scenarios, users often pose questions that extend beyond the informational scope of the video, highlighting the need for Video-LLMs to assess the relevance of the question. We demonstrate that even the best-performing Video-LLMs fail to reject unfit questions-not necessarily due to a lack of video understanding, but because they have not been trained to identify and refuse such questions. To address this limitation, we propose alignment for answerability, a framework that equips Video-LLMs with the ability to evaluate the relevance of a question based on the input video and appropriately decline to answer when the question exceeds the scope of the video, as well as an evaluation framework with a comprehensive set of metrics designed to measure model behavior before and after alignment. Furthermore, we present a pipeline for creating a dataset specifically tailored for alignment for answerability, leveraging existing video-description paired datasets.

488. BrainUICL: An Unsupervised Individual Continual Learning Framework for EEG Applications

链接：https://iclr.cc/virtual/2025/poster/30870 abstract： Electroencephalography (EEG) is a non-invasive brain-computer interface technology used for recording brain electrical activity. It plays an important role in human life and has been widely uesd in real life, including sleep staging, emotion recognition, and motor imagery. However, existing EEG-related models cannot be well applied in practice, especially in clinical settings, where new patients with individual discrepancies appear every day. Such EEG-based model trained on fixed datasets cannot generalize well to the continual flow of numerous unseen subjects in real-world scenarios. This limitation can be addressed through continual learning (CL), wherein the CL model can continuously learn and advance over time. Inspired by CL, we introduce a novel Unsupervised Individual Continual Learning paradigm for handling this issue in practice. We propose the BrainUICL framework, which enables the EEG-based model to continuously adapt to the incoming new subjects. Simultaneously, BrainUICL helps the model absorb new knowledge during each adaptation, thereby advancing its generalization ability for all unseen subjects. The effectiveness of the proposed BrainUICL has been evaluated on three different mainstream EEG tasks. The BrainUICL can effectively balance both the plasticity and stability during CL, achieving better plasticity on new individuals and better stability across all the unseen individuals, which holds significance in a practical setting.

489. Pursuing Better Decision Boundaries for Long-Tailed Object Detection via Category Information Amount

链接：https://iclr.cc/virtual/2025/poster/29994 abstract： In object detection, the number of instances is commonly used to determine whether a dataset follows a long-tailed distribution, implicitly assuming that the model will perform poorly on categories with fewer instances. This assumption has led to extensive research on category bias in datasets with imbalanced instance distributions. However, even in datasets with relatively balanced instance counts, models still exhibit bias toward certain categories, indicating that instance count alone cannot explain this phenomenon. In this work, we first introduce the concept and measurement of category informativeness. We observe a significant negative correlation between a category's informativeness and its accuracy, suggesting that informativeness more accurately reflects the learning difficulty of a category. Based on this observation, we propose the Informativeness-Guided Angular Margin Loss (IGAM Loss), which dynamically adjusts the decision space of categories according to their informativeness, thereby mitigating category bias in long-tailed datasets. IGAM Loss not only achieves superior performance on long-tailed benchmark datasets such as LVIS v1.0 and COCO-LT but also demonstrates significant improvements for underrepresented categories in non-long-tailed datasets like Pascal VOC. Extensive experiments confirm the potential of category informativeness as a tool and the generalizability of our proposed method.

490. Mixture of Experts Made Personalized: Federated Prompt Learning for Vision-Language Models

链接：https://iclr.cc/virtual/2025/poster/27771 abstract： Federated prompt learning benefits federated learning with CLIP-like Vision-Language Model's (VLM's) robust representation learning ability through prompt learning. However, current federated prompt learning methods are habitually restricted to the traditional FL paradigm, where the participating clients are generally only allowed to download a single globally aggregated model from the server. While justifiable for training full-sized models under federated settings, in this work, we argue that this paradigm is ill-suited for lightweight prompts. By facilitating the clients to download multiple pre-aggregated prompts as fixed non-local experts, we propose Personalized Federated Mixture of Adaptive Prompts (pFedMoAP), a novel FL framework that personalizes the prompt learning process through the lens of Mixture of Experts (MoE). pFedMoAP implements a local attention-based gating network that learns to generate enhanced text features for better alignment with local image data, benefiting from both local and downloaded non-local adaptive prompt experts. Extensive experiments on 9 datasets under various federated settings demonstrate the efficacy of the proposed pFedMoAP algorithm. The code is available at https://github.com/ljaiverson/pFedMoAP.

491. Valid Conformal Prediction for Dynamic GNNs

链接：https://iclr.cc/virtual/2025/poster/28722 abstract： Dynamic graphs provide a flexible data abstraction for modelling many sorts of real-world systems, such as transport, trade, and social networks. Graph neural networks (GNNs) are powerful tools allowing for different kinds of prediction and inference on these systems, but getting a handle on uncertainty, especially in dynamic settings, is a challenging problem.In this work we propose to use a dynamic graph representation known in the tensor literature as the unfolding, to achieve valid prediction sets via conformal prediction. This representation, a simple graph, can be input to any standard GNN and does not require any modification to existing GNN architectures or conformal prediction routines. One of our key contributions is a careful mathematical consideration of the different inference scenarios which can arise in a dynamic graph modelling context. For a range of practically relevant cases, we obtain valid prediction sets with almost no assumptions, even dispensing with exchangeability. In a more challenging scenario, which we call the semi-inductive regime, we achieve valid prediction under stronger assumptions, akin to stationarity. We provide real data examples demonstrating validity, showing improved accuracy over baselines, and sign-posting different failure modes which can occur when those assumptions are violated.

492. TestGenEval: A Real World Unit Test Generation and Test Completion Benchmark

链接：https://iclr.cc/virtual/2025/poster/30800 abstract： Code generation models can help improve many common software tasks ranging from code completion to defect prediction. Most of the existing benchmarks for code generation LLMs focus on code authoring or code completion. Surprisingly, there has been far less effort dedicated to benchmarking software testing, despite the strong correlation between well-tested software and effective bug detection. To address this gap, we create and release TestGenEval, a large-scale benchmark to measure test generation performance. Based on SWEBench, TestGenEval comprises 68,647 tests from 1,210 code and test file pairs across 11 well-maintained Python repositories. It covers initial tests authoring, test suite completion, and code coverage improvements. Test authoring simulates the process of a developer writing a test suite from scratch, while test completion mimics the scenario where a developer aims to improve the coverage of an existing test suite. We evaluate several popular models, with sizes ranging from 7B to 405B parameters. Our detailed analysis highlights TestGenEval's contribution to a comprehensive evaluation of test generation performance. In particular, models struggle to generate high-coverage test suites, with the best model, GPT-4o, achieving an average coverage of only 35.2\%. This is primarily due to models struggling to reason about execution, and their frequent assertion errors when addressing complex code paths.

493. LR0.FM: LOW-RESOLUTION ZERO-SHOT CLASSIFICATION BENCHMARK FOR FOUNDATION MODELS

链接：https://iclr.cc/virtual/2025/poster/30609 abstract： Visual-language foundation Models (FMs) exhibit remarkable zero-shot generalization across diverse tasks, largely attributed to extensive pre-training on largescale datasets. However, their robustness on low-resolution/pixelated (LR) images, a common challenge in real-world scenarios, remains underexplored. We introduce LR0.FM, a comprehensive benchmark evaluating the impact of low resolution on the zero-shot classification performance of 10 FM(s) across 66 backbones and 15 datasets. We propose a novel metric, Weighted Aggregated Robustness, to address the limitations of existing metrics and better evaluate model performance across resolutions and datasets. Our key findings show that: (i) model size positively correlates with robustness to resolution degradation, (ii) pre-training dataset quality is more important than its size, and (iii) fine-tuned and higher resolution models are less robust against LR. Our analysis further reveals that the model makes semantically reasonable predictions at LR, and the lack of fine-grained details in input adversely impacts the model’s initial layers more than the deeper layers. We use these insights and introduce a simple strategy, LR-TK0, to enhance the robustness of models without compromising their pre-trained weights. We demonstrate the effectiveness of LR-TK0 for robustness against low-resolution across several datasets and its generalization capability across backbones and other approaches. Code is available at this : https://github.com/shyammarjit/LR0.FM

494. CBraMod: A Criss-Cross Brain Foundation Model for EEG Decoding

链接：https://iclr.cc/virtual/2025/poster/29880 abstract： Electroencephalography (EEG) is a non-invasive technique to measure and record brain electrical activity, widely used in various BCI and healthcare applications. Early EEG decoding methods rely on supervised learning, limited by specific tasks and datasets, hindering model performance and generalizability. With the success of large language models, there is a growing body of studies focusing on EEG foundation models. However, these studies still leave challenges: Firstly, most of existing EEG foundation models employ full EEG modeling strategy. It models the spatial and temporal dependencies between all EEG patches together, but ignores that the spatial and temporal dependencies are heterogeneous due to the unique structural characteristics of EEG signals. Secondly, existing EEG foundation models have limited generalizability on a wide range of downstream BCI tasks due to varying formats of EEG data, making it challenging to adapt to. To address these challenges, we propose a novel foundation model called CBraMod. Specifically, we devise a criss-cross transformer as the backbone to thoroughly leverage the structural characteristics of EEG signals, which can model spatial and temporal dependencies separately through two parallel attention mechanisms. And we utilize an asymmetric conditional positional encoding scheme which can encode positional information of EEG patches and be easily adapted to the EEG with diverse formats. CBraMod is pre-trained on a very large corpus of EEG through patch-based masked EEG reconstruction. We evaluate CBraMod on up to 10 downstream BCI tasks (12 public datasets). CBraMod

achieves the state-of-the-art performance across the wide range of tasks, proving its strong capability and generalizability. The source code is publicly available at https://github.com/wjq-learning/CBraMod.

495. Robust Simulation-Based Inference under Missing Data via Neural Processes

链接：https://iclr.cc/virtual/2025/poster/30254 abstract： Simulation-based inference (SBI) methods typically require fully observed data to infer parameters of models with intractable likelihood functions. However, datasets often contain missing values due to incomplete observations, data corruptions (common in astrophysics), or instrument limitations (e.g., in high-energy physics applications). In such scenarios, missing data must be imputed before applying any SBI method. We formalize the problem of missing data in SBI and demonstrate that naive imputation methods can introduce bias in the estimation of SBI posterior. We also introduce a novel amortized method that addresses this issue by jointly learning the imputation model and the inference network within a neural posterior estimation (NPE) framework. Extensive empirical results on SBI benchmarks show that our approach provides robust inference outcomes compared to standard baselines for varying levels of missing data. Moreover, we demonstrate the merits of our imputation model on two real-world bioactivity datasets (Adrenergic and Kinase assays). Code is available at https://github.com/Aalto-QuML/RISE.

496. NoVo: Norm Voting off Hallucinations with Attention Heads in Large Language Models

链接：https://iclr.cc/virtual/2025/poster/27716 abstract： Hallucinations in Large Language Models (LLMs) remain a major obstacle, particularly in high-stakes applications where factual accuracy is critical. While representation editing and reading methods have made strides in reducing hallucinations, their heavy reliance on specialised tools and training on in-domain samples, makes them difficult to scale and prone to overfitting. This limits their accuracy gains and generalizability to diverse datasets. This paper presents a lightweight method, Norm Voting (NoVo), which harnesses the untapped potential of attention head norms to dramatically enhance factual accuracy in zero-shot multiple-choice questions (MCQs). NoVo begins by automatically selecting truth-correlated head norms with an efficient, inference-only algorithm using only 30 random samples, allowing NoVo to effortlessly scale to diverse datasets. Afterwards, selected head norms are employed in a simple voting algorithm, which yields significant gains in prediction accuracy. On TruthfulQA MC1, NoVo surpasses the current state-of-the-art and all previous methods by an astounding margin---at least 19 accuracy points. NoVo demonstrates exceptional generalization to 20 diverse datasets, with significant gains in over 90\% of them, far exceeding all current representation editing and reading methods. NoVo also reveals promising gains to finetuning strategies and building textual adversarial defence. NoVo's effectiveness with head norms opens new frontiers in LLM interpretability, robustness and reliability. Our code is available at: https://github.com/hozhengyi/novo

497. Temporal Difference Learning: Why It Can Be Fast and How It Will Be Faster

链接：https://iclr.cc/virtual/2025/poster/28658 abstract： Temporal difference (TD) learning represents a fascinating paradox: It is the prime example of a divergent algorithm that has not vanished after its instability was proven. On the contrary, TD continues to thrive in reinforcement learning (RL), suggesting that it provides significant compensatory benefits. Empirical evidence supports this, as many RL tasks require substantial computational resources, and TD delivers a crucial speed advantage that makes these tasks solvable. However, it is limited to cases where the divergence issues are absent or negligible for unknown reasons. So far, the theoretical foundations behind the speed-up are also unclear. In our work, we address these shortcomings of TD by employing techniques for analyzing iterative schemes developed over the past century. Our analysis reveals that TD possesses a mechanism that enables efficient mapping into the smallest eigenspace—an operation previously thought to necessitate costly matrix inversion. Notably, this effect is independent of the conditioning of the problem, making it particularly well-suited for RL tasks characterized by rapidly increasing condition numbers, e.g. through delayed rewards. Our novel theoretical understanding allows us to develop a scalable algorithm that integrates TD’s speed with the reliable convergence of gradient descent (GD). We additionally validate these improvements through a rigorous mathematical proof in two dimensions, as well as experiments on problems where TD and GD falter, providing valuable insights into the future of optimization techniques in artificial intelligence

498. High-quality Text-to-3D Character Generation with SparseCubes and Sparse Transformers.

链接：https://iclr.cc/virtual/2025/poster/28180 abstract： Current state-of-the-art text-to-3D generation methods struggle to produce 3D models with fine details and delicate structures due to limitations in differentiable mesh representation techniques. This limitation is particularly pronounced in anime character generation, where intricate features such as fingers, hair, and facial details are crucial for capturing the essence of the characters.In this paper, we introduce a novel, efficient, sparse differentiable mesh representation method, termed SparseCubes, alongside a sparse transformer network designed to generate high-quality 3D models. Our method significantly reduces computational requirements by over 95% and storage memory by 50%, enabling the creation of higher resolution meshes with enhanced details and delicate structures. We validate the effectiveness of our approach through its application to text-to-3D anime character generation, demonstrating its capability to accurately render

subtle details and thin structures (e.g. individual fingers) in both meshes and textures.

499. E(3)-equivariant models cannot learn chirality: Field-based molecular generation

链接：https://iclr.cc/virtual/2025/poster/28460 abstract： Obtaining the desired effect of drugs is highly dependent on their molecular geometries. Thus, the current prevailing paradigm focuses on 3D point-cloud atom representations, utilizing graph neural network (GNN) parametrizations, with rotational symmetries baked in via E(3) invariant layers. We prove that such models must necessarily disregard chirality, a geometric property of the molecules that cannot be superimposed on their mirror image by rotation and translation. Chirality plays a key role in determining drug safety and potency. To address this glaring issue, we introduce a novel field-based representation, proposing reference rotations that replace rotational symmetry constraints. The proposed model captures all molecular geometries including chirality, while still achieving highly competitive performance with E(3)-based methods across standard benchmarking metrics.

500. ToolDial: Multi-turn Dialogue Generation Method for Tool-Augmented Language Models

链接：https://iclr.cc/virtual/2025/poster/30135 abstract： Tool-Augmented Language Models (TALMs) leverage external APIs to answer user queries across various domains. However, existing benchmark datasets for TALM research often feature simplistic dialogues that do not reflect real-world scenarios, such as the need for models to ask clarifying questions or proactively call additional APIs when essential information is missing. To address these limitations, we construct and release ToolDial, a dataset comprising 11,111 multi-turn dialogues, with an average of 8.95 turns per dialogue, based on APIs from RapidAPI. ToolDial has two key characteristics. First, the dialogues incorporate 16 user and system actions (e.g., request, clarify, fail inform) to capture the rich dynamics of real-world interactions. Second, we simulate dialogues where the system requests necessary information from the user based on API documentation and seeks additional APIs if the user fails to provide the required information. To facilitate this process, we introduce a method for generating an API graph that represents input and output compatibility between APIs. Using ToolDial, we evaluate a suite of language models on their ability to predict correct actions and extract input parameter values for API calls from the dialogue history. Modern language models achieve accuracy scores below 70\%, indicating substantial room for improvement. We provide a detailed analysis of the areas where these models fall short.

501. Deriving Causal Order from Single-Variable Interventions: Guarantees & Algorithm

链接：https://iclr.cc/virtual/2025/poster/27999 abstract： Targeted and uniform interventions to a system are crucial for unveiling causal relationships. While several methods have been developed to leverage interventional data for causal structure learning, their practical application in real-world scenarios often remains challenging. Recent benchmark studies have highlighted these difficulties, even when large numbers of single-variable intervention samples are available. In this work, we demonstrate, both theoretically and empirically, that such datasets contain a wealth of causal information that can be effectively extracted under realistic assumptions about the data distribution. More specifically, we introduce a novel variant of interventional faithfulness, which relies on comparisons between the marginal distributions of each variable across observational and interventional settings, and we introduce a score on causal orders. Under this assumption, we are able to prove strong theoretical guarantees on the optimum of our score that also hold for large-scale settings. To empirically verify our theory, we introduce Intersort, an algorithm designed to infer the causal order from datasets containing large numbers of single-variable interventions by approximately optimizing our score. Intersort outperforms baselines (GIES, DCDI, PC and EASE) on almost all simulated data settings replicating common benchmarks in the field. Our proposed novel approach to modeling interventional datasets thus offers a promising avenue for advancing causal inference, highlighting significant potential for further enhancements under realistic assumptions.

502. Booster: Tackling Harmful Fine-tuning for Large Language Models via Attenuating Harmful Perturbation

链接：https://iclr.cc/virtual/2025/poster/28050 abstract： Harmful fine-tuning attack poses serious safety concerns for large language models' fine-tuning-as-a-service. While existing defenses have been proposed to mitigate the issue, their performances are still far away from satisfactory, and the root cause of the problem has not been fully recovered. To this end, we in this paper show that \textit{harmful perturbation} over the model weights could be a probable cause of alignment-broken. In order to attenuate the negative impact of harmful perturbation, we propose an alignment-stage solution, dubbed Booster. Technically, along with the original alignment loss, we append a loss regularizer in the alignment stage's optimization. The regularizer ensures that the model's harmful loss reduction after the simulated harmful perturbation is attenuated, thereby mitigating the subsequent fine-tuning risk. Empirical results show that Booster can effectively reduce the harmful score of the fine-tuned models while maintaining the performance of downstream tasks. Our code is available at https://github.com/git-disl/Booster

503. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks

链接：https://iclr.cc/virtual/2025/poster/29555 abstract： Embedding models play a crucial role in a variety of downstream tasks, including semantic similarity, information retrieval, and clustering. While there has been a surge of interest in developing universal text embedding models that generalize across tasks (e.g., MTEB), progress in learning universal multimodal embedding models has been comparatively slow, despite their importance and practical applications. In this work, we explore the potential of building universal multimodal embeddings capable of handling a broad range of downstream tasks. Our contributions are twofold: (1) we propose MMEB (Massive Multimodal Embedding Benchmark), which covers four meta-tasks (classification, visual question answering, multimodal retrieval, and visual grounding) and 36 datasets, including 20 training datasets and 16 evaluation datasets spanning both in-distribution and out-of-distribution tasks, and (2) VLM2Vec (Vision-Language Model → Vector), a contrastive training framework that transforms any vision-language model into an embedding model through contrastive training on MMEB. Unlike previous models such as CLIP and BLIP, which encode text and images independently without task-specific guidance, VLM2Vec can process any combination of images and text while incorporating task instructions to generate a fixed-dimensional vector. We develop a series of VLM2Vec models based on state-of-the-art VLMs, including Phi-3.5-V, LLaVA-1.6, and Qwen2-VL, and evaluate them on MMEB's benchmark. With LoRA tuning, VLM2Vec achieves a 10% to 20% improvement over existing multimodal embedding models on MMEB's evaluation sets. Our findings reveal that VLMs are surprisingly strong embedding models.

504. Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and Directionally Aligned Perturbations

链接：https://iclr.cc/virtual/2025/poster/27697 abstract： In this paper, we explore the two-point zeroth-order gradient estimator and identify the distribution of random perturbations that minimizes the estimator's asymptotic variance as the perturbation stepsize tends to zero. We formulate it as a constrained functional optimization problem over the space of perturbation distributions. Our findings reveal that such desired perturbations can align directionally with the true gradient, instead of maintaining a fixed length. While existing research has largely focused on fixed-length perturbations, the potential advantages of directional alignment have been overlooked. To address this gap, we delve into the theoretical and empirical properties of the directionally aligned perturbation (DAP) scheme, which adaptively offers higher accuracy along critical directions. Additionally, we provide a convergence analysis for stochastic gradient descent using δ-unbiased random perturbations, extending existing complexity bounds to a wider range of perturbations. Through empirical evaluations on both synthetic problems and practical tasks, we demonstrate that DAPs outperform traditional methods under specific conditions.

505. Ensembling Diffusion Models via Adaptive Feature Aggregation

链接：https://iclr.cc/virtual/2025/poster/28951 abstract： The success of the text-guided diffusion model has inspired the development and release of numerous powerful diffusion models within the open-source community. These models are typically fine-tuned on various expert datasets, showcasing diverse denoising capabilities. Leveraging multiple high-quality models to produce stronger generation ability is valuable, but has not been extensively studied. Existing methods primarily adopt parameter merging strategies to produce a new static model. However, they overlook the fact that the divergent denoising capabilities of the models may dynamically change across different states, such as when experiencing different prompts, initial noises, denoising steps, and spatial locations. In this paper, we propose a novel ensembling method, Adaptive Feature Aggregation (AFA), which dynamically adjusts the contributions of multiple models at the feature level according to various states (i.e., prompts, initial noises, denoising steps, and spatial locations), thereby keeping the advantages of multiple diffusion models, while suppressing their disadvantages. Specifically, we design a lightweight Spatial-Aware Block-Wise (SABW) feature aggregator that adaptive aggregates the block-wise intermediate features from multiple U-Net denoisers into a unified one. The core idea lies in dynamically producing an individual attention map for each model's features by comprehensively considering various states. It is worth noting that only SABW is trainable with about 50 million parameters, while other models are frozen. Both the quantitative and qualitative experiments demonstrate the effectiveness of our proposed method.

506. Scalable Bayesian Learning with posteriors

链接：https://iclr.cc/virtual/2025/poster/28863 abstract： Although theoretically compelling, Bayesian learning with modern machine learning models is computationally challenging since it requires approximating a high dimensional posterior distribution. In this work, we (i) introduce posteriors, an easily extensible PyTorch library hosting general-purpose implementations making Bayesian learning accessible and scalable to large data and parameter regimes; (ii) present a tempered framing of stochastic gradient Markov chain Monte Carlo, as implemented in posteriors, that transitions seamlessly into optimization and unveils a minor modification to deep ensembles to ensure they are asymptotically unbiased for the Bayesian posterior, and (iii) demonstrate and compare the utility of Bayesian approximations through experiments including an investigation into the cold posterior effect and applications with large language models.posteriors repository: https://github.com/normal-computing/posteriors

507. EVA: Geometric Inverse Design for Fast Protein Motif-Scaffolding with Coupled Flow

链接：https://iclr.cc/virtual/2025/poster/30067 abstract： Motif-scaffolding is a fundamental component of protein design, which aims to construct the scaffold structure that stabilizes motifs conferring desired functions. Recent advances in generative models are promising for designing scaffolds, with two main approaches: training-based and sampling-based methods. Training-based methods are resource-heavy and slow, while training-free sampling-based methods are flexible but require numerous sampling steps and costly, unstable guidance. To speed up and improve sampling-based methods, we analyzed failure cases and found that errors stem from the trade-off between generation and guidance. Thus we proposed to exploit the spatial context and adjust the generative direction to be consistent with guidance to overcome this trade-off. Motivated by this, we formulate motif-scaffolding as a Geometric Inverse Design task inspired by the image inverse problem, and present Evolution-ViA-reconstruction (EVA), a novel sampling-based coupled flow framework on geometric manifolds, which starts with a pretrained flow-based generative model. EVA uses motif-coupled priors to leverage spatial contexts, guiding the generative process along a straighter probability path, with generative directions aligned with guidance in the early sampling steps. EVA is 70× faster than SOTA model RFDiffusion with competitive and even better performance on benchmark tests. Further experiments on real-world cases including vaccine design, multi-motif scaffolding and motif optimal placement searching demonstrate EVA's superior efficiency and effectiveness.

508. Diversity Empowers Intelligence: Integrating Expertise of Software Engineering Agents

链接：https://iclr.cc/virtual/2025/poster/29057 abstract： Large language model (LLM) agents have shown great potential in solving real-world software engineering (SWE) problems. The most advanced open-source SWE agent can resolve over 27% of real GitHub issues in SWE-Bench Lite. However, these sophisticated agent frameworks exhibit varying strengths, excelling in certain tasks while underperforming in others. To fully harness the diversity of these agents, we propose DEI (Diversity Empowered Intelligence), a framework that leverages their unique expertise. DEI functions as a meta-module atop existing SWE agent frameworks, managing agent collectives for enhanced problem-solving. Experimental results show that a DEI-guided committee of agents is able to surpass the best individual agent's performance by a large margin. For instance, a group of open-source SWE agents, with a maximum individual resolve rate of 27.3% on SWE-Bench Lite, can achieve a 34.3% resolve rate with DEI, making a 25% improvement and beating most closed-source solutions. Our best-performing group excels with a 55% resolve rate, securing the highest ranking on SWE-Bench Lite. Our findings contribute to the growing body of research on collaborative AI systems and their potential to solve complex software engineering challenges.

509. Point-SAM: Promptable 3D Segmentation Model for Point Clouds

链接：https://iclr.cc/virtual/2025/poster/27719 abstract： The development of 2D foundation models for image segmentation has been significantly advanced by the Segment Anything Model (SAM). However, achieving similar success in 3D models remains a challenge due to issues such as non-unified data formats, poor model scalability, and the scarcity of labeled data with diverse masks. To this end, we propose a 3D promptable segmentation model Point-SAM, focusing on point clouds. We employ an efficient transformer-based architecture tailored for point clouds, extending SAM to the 3D domain. We then distill the rich knowledge from 2D SAM for Point-SAM training by introducing a data engine to generate part-level and object-level pseudo-labels at scale from 2D SAM. Our model outperforms state-of-the-art 3D segmentation models on several indoor and outdoor benchmarks and demonstrates a variety of applications, such as interactive 3D annotation and zero-shot 3D instance proposal.

510. Sharper Guarantees for Learning Neural Network Classifiers with Gradient Methods

链接：https://iclr.cc/virtual/2025/poster/28778 abstract： In this paper, we study the data-dependent convergence and generalization behavior of gradient methods for neural networks with smooth activation. Our first result is a novel bound on the excess risk of deep networks trained by the logistic loss via an alogirthmic stability analysis. Compared to previous works, our results improve upon the shortcomings of the well-established Rademacher complexity-based bounds. Importantly, the bounds we derive in this paper are tighter, hold even for neural networks of small width, do not scale unfavorably with width, are algorithm-dependent, and consequently capture the role of initialization on the sample complexity of gradient descent for deep nets. Specialized to noiseless data separable with margin γ by neural tangent kernel (NTK) features of a network of width $\Omega(poly(\log(n)))$, we show the test-error rate $e^{O(L)}/{\gamma^2 n}$, where n is the training set size and L denotes the number of hidden layers. This results in an improvement in the test loss bound compared to previous works while maintaining the poly-logarithmic width conditions. We further investigate excess risk bounds for deep nets trained with noisy data, establishing that under a polynomial condition on the network width, gradient descent can achieve the optimal excess risk. Finally, we show that a large step-size significantly improves upon the NTK regime's results in classifying the XOR distribution. In particular, we show for a one-hidden layer neural network of constant width m with quadratic activation and standard Gaussian initialization that SGD with linear sample complexity and with a large step-size $\eta=m$ reaches the perfect test accuracy after only $\lceil\log(d)\rceil$ iterations, where d is the data dimension.

511. TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models

链接：https://iclr.cc/virtual/2025/poster/29025 abstract： Causal language models have demonstrated remarkable capabilities, but their size poses significant challenges for deployment in resource-constrained environments. Knowledge distillation, a widely-used technique for transferring knowledge from a large teacher model to a small student model, presents a promising approach for model compression.A significant remaining issue lies in the major differences between teacher and student models, namely the substantial capacity gap, mode averaging, and mode collapse, which pose barriers during distillation.To address these issues, we introduce $\textit{Temporally Adaptive Interpolated Distillation (TAID)}$, a novel knowledge distillation approach that dynamically interpolates student and teacher distributions through an adaptive intermediate distribution, gradually shifting from the student's initial distribution towards the teacher's distribution. We provide a theoretical analysis demonstrating TAID's ability to prevent mode collapse and empirically show its effectiveness in addressing the capacity gap while balancing mode averaging and mode collapse.Our comprehensive experiments demonstrate TAID's superior performance across various model sizes and architectures in both instruction tuning and pre-training scenarios. Furthermore, we showcase TAID's practical impact by developing two state-of-the-art compact foundation models: $\texttt{TAID-LLM-1.5B}$ for language tasks and $\texttt{TAID-VLM-2B}$ for vision-language tasks.These results demonstrate TAID's effectiveness in creating high-performing and efficient models, advancing the development of more accessible AI technologies.

512. SoftMatcha: A Soft and Fast Pattern Matcher for Billion-Scale Corpus Searches

链接：https://iclr.cc/virtual/2025/poster/29709 abstract： Researchers and practitioners in natural language processing and computational linguistics frequently observe and analyze the real language usage in large-scale corpora.For that purpose, they often employ off-the-shelf pattern-matching tools, such as grep, and keyword-in-context concordancers, which is widely used in corpus linguistics for gathering examples.Nonetheless, these existing techniques rely on surface-level string matching, and thus they suffer from the major limitation of not being able to handle orthographic variations and paraphrasing---notable and common phenomena in any natural language.In addition, existing continuous approaches such as dense vector search tend to be overly coarse, often retrieving texts that are unrelated but share similar topics.Given these challenges, we propose a novel algorithm that achieves soft (or semantic) yet efficient pattern matching by relaxing a surface-level matching with word embeddings.Our algorithm is highly scalable with respect to the size of the corpus text utilizing inverted indexes.We have prepared an efficient implementation, and we provide an accessible web tool.Our experiments demonstrate that the proposed method(i) can execute searches on billion-scale corpora in less than a second, which is comparable in speed to surface-level string matching and dense vector search;(ii) can extract harmful instances that semantically match queries from a large set of English and Japanese Wikipedia articles;and (iii) can be effectively applied to corpus-linguistic analyses of Latin, a language with highly diverse inflections.

513. LICORICE: Label-Efficient Concept-Based Interpretable Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/29920 abstract： Recent advances in reinforcement learning (RL) have predominantly leveraged neural network policies for decision-making, yet these models often lack interpretability, posing challenges for stakeholder comprehension and trust. Concept bottleneck models offer an interpretable alternative by integrating human-understandable concepts into policies. However, prior work assumes that concept annotations are readily available during training. For RL, this requirement poses a significant limitation: it necessitates continuous real-time concept annotation, which either places an impractical burden on human annotators or incurs substantial costs in API queries and inference time when employing automated labeling methods. To overcome this limitation, we introduce a novel training scheme that enables RL agents to efficiently learn a concept-based policy by only querying annotators to label a small set of data. Our algorithm, LICORICE, involves three main contributions: interleaving concept learning and RL training, using an ensemble to actively select informative data points for labeling, and decorrelating the concept data. We show how LICORICE reduces human labeling efforts to 500 or fewer concept labels in three environments, and 5000 or fewer in two more complex environments, all at no cost to performance. We also explore the use of VLMs as automated concept annotators, finding them effective in some cases but imperfect in others. Our work significantly reduces the annotation burden for interpretable RL, making it more practical for real-world applications that necessitate transparency. Our code is released.

514. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second

链接：https://iclr.cc/virtual/2025/poster/29138 abstract： We present a foundation model for zero-shot metric monocular depth estimation. Our model, Depth Pro, synthesizes high-resolution depth maps with unparalleled sharpness and high-frequency details. The predictions are metric, with absolute scale, without relying on the availability of metadata such as camera intrinsics. And the model is fast, producing a 2.25-megapixel depth map in 0.3 seconds on a standard GPU. These characteristics are enabled by a number of technical contributions, including an efficient multi-scale vision transformer for dense prediction, a training protocol that combines real and synthetic datasets to achieve high metric accuracy alongside fine boundary tracing, dedicated evaluation metrics for boundary accuracy in estimated depth maps, and state-of-the-art focal length estimation from a single image. Extensive experiments analyze specific design choices and demonstrate that Depth Pro outperforms prior work along multiple dimensions. We release code & weights at https://github.com/apple/ml-depth-pro

515. Cross-Modal Safety Mechanism Transfer in Large Vision-Language

Models

链接：https://iclr.cc/virtual/2025/poster/31036 abstract： Vision-language alignment in Large Vision-Language Models (LVLMs) successfully enables LLMs to understand visual input. However, we find that existing vision-language alignment methods fail to transfer the existing safety mechanism for text in LLMs to vision, which leads to vulnerabilities in toxic image. To explore the cause of this problem, we give the insightful explanation of where and how the safety mechanism of LVLMs operates and conduct comparative analysis between text and vision. We find that the hidden states at the specific transformer layers play a crucial role in the successful activation of safety mechanism, while the vision-language alignment at hidden states level in current methods is insufficient. This results in a semantic shift for input images compared to text in hidden states, therefore misleads the safety mechanism. To address this, we propose a novel Text-Guided vision-language Alignment method (TGA) for LVLMs. TGA retrieves the texts related to input vision and uses them to guide the projection of vision into the hidden states space in LLMs. Experiments show that \textbf{TGA} not only successfully transfers the safety mechanism for text in basic LLMs to vision in vision-language alignment for LVLMs without any safety fine-tuning on the visual modality but also maintains the general performance on various vision tasks (Safe and Good). Code is in supplemental material and will be released on GitHub after acceptance.

516. Video-STaR: Self-Training Enables Video Instruction Tuning with Any Supervision

链接：https://iclr.cc/virtual/2025/poster/30101 abstract： The performance and reasoning capabilities of Large Multi-modal Models (LMMs) is dependent on the size and quality of their training datasets. However, collecting datasets that support chain-of-thought instruction tuning is highly challenging. Existing video instruction tuning datasets are often derived by prompting large language models with video captions to generate question-answer pairs, which makes them predominantly descriptive rather than reasoning-focused. Meanwhile, many labeled video datasets with diverse labels and supervision exist -- however, we find that their integration into LMMs is non-trivial. Herein, we present $\underline{\text{Video}}$ $\underline{\text{S}}\text{elf}$-$\underline{\text{T}}\text{raining}$ with $\underline{\text{a}}\text{ugmented}$ $\underline{\text{R}}\text{easoning}$ (Video-STaR), the first self-training approach for video instruction tuning. Video-STaR allows the utilization of *any* labeled video dataset for video instruction tuning.In Video-STaR, an LMM cycles between instruction generation and finetuning, which we show (I) improves general video understanding and (II) adapts LMMs to novel downstream tasks with existing supervision. During instruction generation, an LMM is prompted to propose an answer. The answers are then filtered only to those that contain the original video labels, and the LMM is then re-trained on the generated dataset. By training exclusively on generated answers containing the correct video labels, Video-STaR leverages these existing labels as weak supervision for video instruction tuning.Our results demonstrate that Video-STaR-augmented LMMs achieve notable improvements in (I) general Video QA, where TempCompass performance improved by 6.1%, *and* (II) downstream tasks, with a 9.9% increase in Kinetics700-QA accuracy and a 4.0% improvement in action quality assessment on FineDiving, while also exhibiting better interpretability.

517. COFlowNet: Conservative Constraints on Flows Enable High-Quality Candidate Generation

链接：https://iclr.cc/virtual/2025/poster/28047 abstract： Generative flow networks (GFlowNets) have been considered as powerful tools for generating candidates with desired properties. Given that evaluating the property of candidates can be complex and time-consuming, existing GFlowNets train proxy models for efficient online evaluation. However, the performance of proxy models is heavily dependent on the amount of data and is of considerable uncertainty. Therefore, it is of great interest that how to develop an offline GFlowNet that does not rely on online evaluation. Under the offline setting, the limited data results in an insufficient exploration of state space. The insufficient exploration means that offline GFlowNets can hardly generate satisfying candidates out of the distribution of training data. Therefore, it is critical to restrict the offline model to act in the distribution of training data. The distinctive training goal of GFlownets poses a unique challenge for making such restrictions. Tackling the challenge, we propose Conservative Offline GFlowNet (COFlowNet) in this paper. We define unsupported flow, edges containing unseen states in training data. Models can learn extremely little knowledge about unsupported flow from training data. By constraining the model from exploring unsupported flows, we restrict COFlowNet to explore as optimal trajectories on the training set as possible, thus generating better candidates. In order to improve the diversity of candidates, we further introduce a quantile version of unsupported flow restriction. Experimental results on several widely-used datasets validate the effectiveness of COFlowNet in generating high-scored and diverse candidates. All implementations are available at https://github.com/yuxuan9982/COflownet.

518. Training-free Camera Control for Video Generation

链接：https://iclr.cc/virtual/2025/poster/30066 abstract： We propose a training-free and robust solution to offer camera movement control for off-the-shelf video diffusion models. Unlike previous work, our method does not require any supervised finetuning on camera-annotated datasets or self-supervised training via data augmentation. Instead, it is plug-and-play with most pretrained video diffusion models and can generate camera-controllable videos with a single image or text prompt as input. The inspiration for our work comes from the layout prior that intermediate latents encode for the generated results, thus rearranging noisy pixels in them will cause the output content to relocate as well. As camera moving could also be seen as a type of pixel rearrangement caused by perspective change, videos can be reorganized following specific camera motion if their noisy latents change accordingly. Building on this, we propose CamTrol, which enables robust camera control for video diffusion models. It is

achieved by a two-stage process. First, we model image layout rearrangement through explicit camera movement in 3D point cloud space. Second, we generate videos with camera motion by leveraging the layout prior of noisy latents formed by a series of rearranged images. Extensive experiments have demonstrated its superior performance in both video generation and camera motion alignment compared with other finetuned methods. Furthermore, we show the capability of CamTrol to generalize to various base models, as well as its impressive applications in scalable motion control, dealing with complicated trajectories and unsupervised 3D video generation. Videos available at https://lifedecoder.github.io/CamTrol/.

519. LongMamba: Enhancing Mamba's Long-Context Capabilities via Training-Free Receptive Field Enlargement

链接：https://iclr.cc/virtual/2025/poster/28881 abstract： State space models (SSMs) have emerged as an efficient alternative to Transformer models for language modeling, offering linear computational complexity and constant memory usage as context length increases. However, despite their efficiency in handling long contexts, recent studies have shown that SSMs, such as Mamba models, generally underperform compared to Transformers in long-context understanding tasks. To address this significant shortfall and achieve both efficient and accurate long-context understanding, we propose LongMamba, a training-free technique that significantly enhances the long-context capabilities of Mamba models. LongMamba builds on our discovery that the hidden channels in Mamba can be categorized into local and global channels based on their receptive field lengths, with global channels primarily responsible for long-context capability. These global channels can become the key bottleneck as the input context lengthens. Specifically, when input lengths largely exceed the training sequence length, global channels exhibit limitations in adaptively extend their receptive fields, leading to Mamba’s poor long-context performance. The key idea of LongMamba is to mitigate the hidden state memory decay in these global channels by preventing the accumulation of unimportant tokens in their memory. This is achieved by first identifying critical tokens in the global channels and then applying token filtering to accumulate only those critical tokens. Through extensive benchmarking across synthetic and real-world long-context scenarios, LongMamba sets a new standard for Mamba’s long-context performance, significantly extending its operational range without requiring additional training. Our code is available at https://github.com/GATECH-EIC/LongMamba.

520. Learning Neural Networks with Distribution Shift: Efficiently Certifiable Guarantees

链接：https://iclr.cc/virtual/2025/poster/28915 abstract： We give the first provably efficient algorithms for learning neural networks with respect to distribution shift. We work in the Testable Learning with Distribution Shift framework (TDS learning) of Klivans et al. (2024), where the learner receives labeled examples from a training distribution and unlabeled examples from a test distribution and must either output a hypothesis with low test error or reject if distribution shift is detected. No assumptions are made on the test distribution. All prior work in TDS learning focuses on classification, while here we must handle the setting of nonconvex regression. Our results apply to real-valued networks with arbitrary Lipschitz activations and work whenever the training distribution has strictly sub-exponential tails. For training distributions that are bounded and hypercontractive, we give a fully polynomial-time algorithm for TDS learning one hidden-layer networks with sigmoid activations. We achieve this by importing classical kernel methods into the TDS framework using data-dependent feature maps and a type of kernel matrix that couples samples from both train and test distributions.

521. Adaptive backtracking for faster optimization

链接：https://iclr.cc/virtual/2025/poster/29575 abstract： Backtracking line search is foundational in numerical optimization. The basic idea is to adjust the step size of an algorithm by a {\em constant} factor until some chosen criterion (e.g. Armijo, Descent Lemma) is satisfied. We propose a novel way to adjust step sizes, replacing the constant factor used in regular backtracking with one that takes into account the degree to which the chosen criterion is violated, with no additional computational burden. This light-weight adjustment leads to significantly faster optimization, which we confirm by performing a variety of experiments on over fifteen real world datasets.For convex problems, we prove adaptive backtracking requires no more adjustments to produce a feasible step size than regular backtracking does.For nonconvex smooth problems, we prove adaptive backtracking enjoys the same guarantees of regular backtracking. %same lower bounds that step sizes produced by regular backtracking do.Furthermore, we prove adaptive backtracking preserves the convergence rates of gradient descent and its accelerated variant.

522. CodePlan: Unlocking Reasoning Potential in Large Language Models by Scaling Code-form Planning

链接：https://iclr.cc/virtual/2025/poster/29000 abstract： Despite the remarkable success of large language models (LLMs) on traditional natural language processing tasks, their planning ability remains a critical bottleneck in tackling complex multi-step reasoning tasks. Existing approaches mainly rely on prompting or task-specific fine-tuning, often suffering from weak robustness and cross-task generalization. To address the limitation, we introduce CodePlan, a scalable paradigm that empowers LLMs to generate and follow code-form plans---pseudocode that outlines high-level, structured reasoning processes. By leveraging the structured and versatile nature of code, CodePlan effectively captures the rich semantics and control flows inherent to sophisticated reasoning. Importantly, CodePlan allows the automatic extraction of code-form plans from massive, wide-ranging text corpora without the need for curated, task-specific datasets. This enables it to scale up efficiently and improve reasoning

capabilities across diverse scenarios. To train CodePlan, we construct a large-scale dataset of 2M examples that integrate code-form plans with standard prompt-response pairs from existing corpora. With minimal computation overhead during both training and inference, CodePlan achieves a 25.1\% relative improvement compared with directly generating responses, averaged across 13 challenging multi-step reasoning benchmarks, spanning mathematical reasoning, symbolic reasoning, instruction-following, multi-hop QA, and decision-making tasks. Further analysis reveals CodePlan's increasing performance gains on more complex reasoning tasks, as well as significant data efficiency thanks to its generalization ability.

523. Accelerated training through iterative gradient propagation along the residual path

链接：https://iclr.cc/virtual/2025/poster/30119 abstract： Despite being the cornerstone of deep learning, backpropagation is criticized for its inherent sequentiality, which can limit the scalability of very deep models.Such models faced convergence issues due to vanishing gradient, later resolved using residual connections. Variants of these are now widely used in modern architectures.However, the computational cost of backpropagation remains a major burden, accounting for most of the training time.Taking advantage of residual-like architectural designs, we introduce Highway backpropagation, a parallelizable iterative algorithm that approximates backpropagation, by alternatively i) accumulating the gradient estimates along the residual path, and ii) backpropagating them through every layer in parallel. This algorithm is naturally derived from a decomposition of the gradient as the sum of gradients flowing through all paths, and is adaptable to a diverse set of common architectures, ranging from ResNets and Transformers to recurrent neural networks.Through an extensive empirical study on a large selection of tasks and models, we evaluate Highway-BP and show that major speedups can be achieved with minimal performance degradation.

524. Continuous Exposure Learning for Low-light Image Enhancement using Neural ODEs

链接：https://iclr.cc/virtual/2025/poster/29919 abstract： Low-light image enhancement poses a significant challenge due to the limited information captured by image sensors in low-light environments. Despite recent improvements in deep learning models, the lack of paired training datasets remains a significant obstacle. Therefore, unsupervised methods have emerged as a promising solution. In this work, we focus on the strength of curve-adjustment-based approaches to tackle unsupervised methods. The majority of existing unsupervised curve-adjustment approaches iteratively estimate higher order curve parameters to enhance the exposure of images while efficiently preserving the details of the images. However, the convergence of the enhancement procedure cannot be guaranteed, leading to sensitivity to the number of iterations and limited performance. To address this problem, we consider the iterative curve-adjustment update process as a dynamic system and formulate it as a Neural Ordinary Differential Equations (NODE) for the first time, and this allows us to learn a continuous dynamics of the latent image. The strategy of utilizing NODE to leverage continuous dynamics in iterative methods enhances unsupervised learning and aids in achieving better convergence compared to discrete-space approaches. Consequently, we achieve state-of-the-art performance in unsupervised low-light image enhancement across various benchmark datasets.

525. JetFormer: An autoregressive generative model of raw images and text

链接：https://iclr.cc/virtual/2025/poster/28106 abstract： Removing modeling constraints and unifying architectures across domains has been a key driver of the recent progress in training large multimodal models. However, most of these models still rely on many separately trained components such as modality-specific encoders and decoders. In this work, we further streamline joint generative modeling of images and text. We propose an autoregressive decoder-only transformer---JetFormer---which is trained to directly maximize the likelihood of raw data, without relying on any separately pretrained components, and can understand and generate both text and images. Specifically, we leverage a normalizing flow model to obtain a soft-token image representation that is jointly trained with an autoregressive multimodal transformer. The normalizing flow model serves as both an image encoder for perception tasks and an image decoder for image generation tasks during inference. JetFormer achieves text-to-image generation quality competitive with recent VQVAE- and VAE-based baselines. These baselines rely on pretrained image autoencoders, which are trained with a complex mixture of losses, including perceptual ones. At the same time, JetFormer demonstrates robust image understanding capabilities. To the best of our knowledge, JetFormer is the first model that is capable of generating high-fidelity images and producing strong log-likelihood bounds.

526. Boosting Perturbed Gradient Ascent for Last-Iterate Convergence in Games

链接：https://iclr.cc/virtual/2025/poster/30088 abstract： This paper presents a payoff perturbation technique, introducing a strong convexity to players' payoff functions in games. This technique is specifically designed for first-order methods to achieve last-iterate convergence in games where the gradient of the payoff functions is monotone in the strategy profile space, potentially containing additive noise. Although perturbation is known to facilitate the convergence of learning algorithms, the magnitude of perturbation requires careful adjustment to ensure last-iterate convergence. Previous studies have proposed a scheme in which the magnitude is determined by the distance from a periodically re-initialized anchoring or reference strategy. Building upon this, we propose Gradient Ascent with Boosting Payoff Perturbation, which incorporates a novel perturbation into the underlying payoff function, maintaining the periodically re-initializing anchoring strategy scheme. This innovation empowers us to provide faster last-iterate convergence rates against the existing payoff perturbed algorithms, even in the presence of additive noise.

527. AdaRankGrad: Adaptive Gradient Rank and Moments for Memory-Efficient LLMs Training and Fine-Tuning

链接：https://iclr.cc/virtual/2025/poster/29966 abstract： Training and fine-tuning large language models (LLMs) come with challenges related to memory and computational requirements due to the increasing size of the model weights and the optimizer states. To tackle these challenges, various techniques have been developed, such as low-rank adaptation (LoRA), which involves introducing a parallel trainable low-rank matrix to the fixed pre-trained weights at each layer. However, these methods often fall short compared to the full-rank weight training approach, as they restrict the parameter search to a low-rank subspace. This limitation can disrupt training dynamics and may require a full-rank warm start to mitigate the impact. In this paper, we introduce a new method inspired by a phenomenon we formally prove: as training progresses, the rank of the estimated layer gradients gradually decreases and asymptotically approaches rank one. Leveraging this, our approach involves adaptively reducing the rank of the gradients during Adam optimization steps, using an efficient online-updating low-rank projections rule. We further present a randomized-svd scheme for efficiently finding the projection matrix. Our technique enables full-parameter fine-tuning with adaptive low-rank gradient updates, significantly reducing overall memory requirements during training compared to state-of-the-art methods while improving model performance in both pretraining and fine-tuning. Finally, we provide a convergence analysis of our method and demonstrate its merits for training and fine-tuning language and biological foundation models.

528. Towards a Theoretical Understanding of Synthetic Data in LLM Post-Training: A Reverse-Bottleneck Perspective

链接：https://iclr.cc/virtual/2025/poster/29438 abstract： Synthetic data has become a pivotal resource in post-training tasks for large language models (LLMs) due to the scarcity of high-quality, specific data. While various methods have been developed to generate synthetic data, there remains a discernible gap between the practical effects of synthetic data and our theoretical comprehension. To address this challenge, we commence by presenting a detailed modeling of the prevalent synthetic data generation process. Building upon this modeling, we demonstrate that the generalization capability of the post-trained model is critically determined by the information gain derived from the generative model, as analyzed from a novel reverse-bottleneck perspective. Moreover, we introduce the concept of Generalization Gain via Mutual Information (GGMI) and elucidate the relationship between generalization gain and information gain. This analysis serves as a theoretical foundation for synthetic data generation and further highlights its connection with the generalization capability of post-trained models, offering an understanding about the design of synthetic data generation techniques and the optimization of the post-training process. We open-source our code at https://github.com/ZyGan1999/Towards-a-Theoretical-Understanding-of-Synthetic-Data-in-LLM-Post-Training.

529. Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors

链接：https://iclr.cc/virtual/2025/poster/29637 abstract： This paper studies how to synthesize face images of non-existent persons, to create a dataset that allows effective training of face recognition (FR) models. Besides generating realistic face images, two other important goals are: 1) the ability to generate a large number of distinct identities (inter-class separation), and 2) a proper variation in appearance of the images for each identity (intra-class variation).However, existing works 1) are typically limited in how many well-separated identities can be generated and 2) either neglect or use an external model for attribute augmentation. We propose Vec2Face, a holistic model that uses only a sampled vector as input and can flexibly generate and control the identity of face images and their attributes. Composed of a feature masked autoencoder and an image decoder, Vec2Face is supervised by face image reconstruction and can be conveniently used in inference. Using vectors with low similarity among themselves as inputs, Vec2Face generates well-separated identities. Randomly perturbing an input identity vector within a small range allows Vec2Face to generate faces of the same identity with proper variation in face attributes. It is also possible to generate images with designated attributes by adjusting vector values with a gradient descent method. Vec2Face has efficiently synthesized as many as 300K identities, whereas 60K is the largest number of identities created in the previous works. As for performance, FR models trained with the generated HSFace datasets, from 10k to 300k identities, achieve state-of-the-art accuracy, from 92\% to 93.52\%, on five real-world test sets (\emph{i.e.}, LFW, CFP-FP, AgeDB-30, CALFW, and CPLFW). For the first time, the FR model trained using our synthetic training set achieves higher accuracy than that trained using a same-scale training set of real face images on the CALFW, IJBB, and IJBC test sets.

530. Lumina-T2X: Scalable Flow-based Large Diffusion Transformer for Flexible Resolution Generation

链接：https://iclr.cc/virtual/2025/poster/30398 abstract： Sora unveils the potential of scaling Diffusion Transformer (DiT) for generating photorealistic images and videos at arbitrary resolutions, aspect ratios, and durations, yet it still lacks sufficient implementation details. In this paper, we introduce the Lumina-T2X family -- a series of Flow-based Large Diffusion Transformers (Flag-DiT) equipped with zero-initialized attention, as a simple and scalable generative framework that can be adapted to various modalities, e.g., transforming noise into images, videos, multi-view 3D objects, or audio clips conditioned on text instructions. By tokenizing the latent spatial-temporal space and incorporating learnable placeholders such as |[nextline]| and |[nextframe]| tokens, Lumina-T2X seamlessly unifies the representations of different modalities across various spatial-temporal

resolutions. Advanced techniques like RoPE, KQ-Norm, and flow matching enhance the stability, flexibility, and scalability of Flag-DiT, enabling models of Lumina-T2X to scale up to 7 billion parameters and extend the context window to 128K tokens. This is particularly beneficial for creating ultra-high-definition images with our Lumina-T2I model and long 720p videos with our Lumina-T2V model. Remarkably, Lumina-T2I, powered by a 5-billion-parameter Flag-DiT, requires only 35% of the training computational costs of a 600-million-parameter naive DiT (PixArt-alpha), indicating that increasing the number of parameters significantly accelerates convergence of generative models without compromising visual quality. Our further comprehensive analysis underscores Lumina-T2X's preliminary capability in resolution extrapolation, high-resolution editing, generating consistent 3D views, and synthesizing videos with seamless transitions. All code and checkpoints of Lumina-T2X are released at https://github.com/Alpha-VLLM/Lumina-T2X to further foster creativity, transparency, and diversity in the generative AI community.

531. Adjoint Matching: Fine-tuning Flow and Diffusion Generative Models with Memoryless Stochastic Optimal Control

链接：https://iclr.cc/virtual/2025/poster/27782 abstract： Dynamical generative models that produce samples through an iterative process, such as Flow Matching and denoising diffusion models, have seen widespread use, but there have not been many theoretically-sound methods for improving these models with reward fine-tuning. In this work, we cast reward fine-tuning as stochastic optimal control (SOC). Critically, we prove that a very specific memoryless noise schedule must be enforced during fine-tuning, in order to account for the dependency between the noise variable and the generated samples. We also propose a new algorithm named Adjoint Matching which outperforms existing SOC algorithms, by casting SOC problems as a regression problem. We find that our approach significantly improves over existing methods for reward fine-tuning, achieving better consistency, realism, and generalization to unseen human preference reward models, while retaining sample diversity.

532. REVISITING MULTI-PERMUTATION EQUIVARIANCE THROUGH THE LENS OF IRREDUCIBLE REPRESENTATIONS

链接：https://iclr.cc/virtual/2025/poster/30985 abstract： This paper explores the characterization of equivariant linear layers for representations of permutations and related groups. Unlike traditional approaches,which address these problems using parameter-sharing, we consider an alternativemethodology based on irreducible representations and Schur’s lemma. Using thismethodology, we obtain an alternative derivation for existing models like DeepSets,2-IGN graph equivariant networks, and Deep Weight Space (DWS) networks. Thederivation for DWS networks is significantly simpler than that of previous results.Next, we extend our approach to unaligned symmetric sets, where equivarianceto the wreath product of groups is required. Previous works have addressed thisproblem in a rather restrictive setting, in which almost all wreath equivariant layersare Siamese. In contrast, we give a full characterization of layers in this case andshow that there is a vast number of additional non-Siamese layers in some settings.We also show empirically that these additional non-Siamese layers can improveperformance in tasks like graph anomaly detection, weight space alignment, andlearning Wasserstein distances.

533. Targeted Attack Improves Protection against Unauthorized Diffusion Customization

链接：https://iclr.cc/virtual/2025/poster/29155 abstract： Diffusion models build a new milestone for image generation yet raising public concerns, for they can be fine-tuned on unauthorized images for customization. Protection based on adversarial attacks rises to encounter this unauthorized diffusion customization, by adding protective watermarks to images and poisoning diffusion models. However, current protection, leveraging untargeted attacks, does not appear to be effective enough. In this paper, we propose a simple yet effective improvement for the protection against unauthorized diffusion customization by introducing targeted attacks. We show that by carefully selecting the target, targeted attacks significantly outperform untargeted attacks in poisoning diffusion models and degrading the customization image quality. Extensive experiments validate the superiority of our method on two mainstream customization methods of diffusion models, compared to existing protections. To explain the surprising success of targeted attacks, we delve into the mechanism of attack-based protections and propose a hypothesis based on our observation, which enhances the comprehension of attack-based protections. To the best of our knowledge, we are the first to both reveal the vulnerability of diffusion models to targeted attacks and leverage targeted attacks to enhance protection against unauthorized diffusion customization.

534. Quest: Query-centric Data Synthesis Approach for Long-context Scaling of Large Language Model

链接：https://iclr.cc/virtual/2025/poster/28141 abstract： Recent advancements in large language models (LLMs) have highlighted the importance of extending context lengths for handling complex tasks. While traditional methods for training on long contexts often use filtered long documents, these approaches lead to domain imbalances, limiting model performance. To address this, techniques like random document concatenation (Standard) and similarity-based methods (KNN, ICLM) have been developed. However, they either sacrifice semantic coherence or diversity. To balance both aspects, we introduce Quest, a query-centric data synthesis method aggregating semantically relevant yet diverse documents. Quest uses a generative model to predict potential queries for each document, grouping documents with similar queries and keywords. Extensive experiments demonstrate Quest's superior performance on long-context tasks, achieving remarkable results with context lengths of up to 1M

tokens and confirming its scalability across various model sizes.

535. On the Expressive Power of Sparse Geometric MPNNs

链接：https://iclr.cc/virtual/2025/poster/29870 abstract： Motivated by applications in chemistry and other sciences, we study the expressivepower of message-passing neural networks for geometric graphs, whose nodefeatures correspond to 3-dimensional positions. Recent work has shown that suchmodels can separate generic pairs of non-isomorphic geometric graphs, though theymay fail to separate some rare and complicated instances. However, these resultsassume a fully connected graph, where each node possesses complete knowledgeof all other nodes. In contrast, often, in application, every node only possessesknowledge of a small number of nearest neighbors.This paper shows that generic pairs of non-isomorphic geometric graphs canbe separated by message-passing networks with rotation equivariant features aslong as the underlying graph is connected. When only invariant intermediatefeatures are allowed, generic separation is guaranteed for generically globallyrigid graphs. We introduce a simple architecture, EGENNET, which achieves ourtheoretical guarantees and compares favorably with alternative architecture onsynthetic and chemical benchmarks

536. Radar: Fast Long-Context Decoding for Any Transformer

链接：https://iclr.cc/virtual/2025/poster/29218 abstract： Transformer models have demonstrated exceptional performance across a wide range of applications. Though forming the foundation of Transformer models, the dot-product attention does not scale well to long-context data since its time requirement grows quadratically with context length. In this work, we propose Radar, a training-free approach that accelerates inference by dynamically searching for the most important context tokens. For any pre-trained Transformer, Radar can reduce the decoding time complexity without training or heuristically evicting tokens. Moreover, we provide theoretical justification for our approach, demonstrating that Radar can reliably identify the most important tokens with high probability. We conduct extensive comparisons with the previous methods on a wide range of tasks. The results demonstrate that Radar achieves the state-of-the-art performance across different architectures with reduced time complexity, offering a practical solution for efficient long-context processing of Transformers. The code is publicly available at https://github.com/BorealisAI/radar-decoding.

537. Credal Wrapper of Model Averaging for Uncertainty Estimation in Classification

链接：https://iclr.cc/virtual/2025/poster/29022 abstract： This paper presents an innovative approach, called credal wrapper, to formulating a credal set representation of model averaging for Bayesian neural networks (BNNs) and deep ensembles (DEs), capable of improving uncertainty estimation in classification tasks. Given a finite collection of single predictive distributions derived from BNNs or DEs, the proposed credal wrapper approach extracts an upper and a lower probability bound per class, acknowledging the epistemic uncertainty due to the availability of a limited amount of distributions. Such probability intervals over classes can be mapped on a convex set of probabilities (a credal set) from which, in turn, a unique prediction can be obtained using a transformation called intersection probability transformation. In this article, we conduct extensive experiments on several out-of-distribution (OOD) detection benchmarks, encompassing various dataset pairs (CIFAR10/100 vs SVHN/Tiny-ImageNet, CIFAR10 vs CIFAR10-C, CIFAR100 vs CIFAR100-C and ImageNet vs ImageNet-O) and using different network architectures (such as VGG16, ResNet-18/50, EfficientNet B2, and ViT Base). Compared to the BNN and DE baselines, the proposed credal wrapper method exhibits superior performance in uncertainty estimation and achieves a lower expected calibration error on corrupted data.

538. An Asynchronous Bundle Method for Distributed Learning Problems

链接：https://iclr.cc/virtual/2025/poster/30025 abstract： We propose a novel asynchronous bundle method to solve distributed learning problems. Compared to existing asynchronous methods, our algorithm computes the next iterate based on a more accurate approximation of the objective function and does not require any prior information about the maximal information delay in the system. This makes the proposed method fast and easy to tune. We prove that the algorithm converges in both deterministic and stochastic (mini-batch) settings, and quantify how the convergence times depend on the level of asynchrony. The practical advantages of our method are illustrated through numerical experiments on classification problems of varying complexities and scales.

539. Learning the Optimal Stopping for Early Classification within Finite Horizons via Sequential Probability Ratio Test

链接：https://iclr.cc/virtual/2025/poster/29607 abstract： Time-sensitive machine learning benefits from Sequential Probability Ratio Test (SPRT), which provides an optimal stopping time for early classification of time series. However, in finite horizon scenarios, where input lengths are finite, determining the optimal stopping rule becomes computationally intensive due to the need for backward induction, limiting practical applicability. We thus introduce FIRMBOUND, an SPRT-based framework that efficiently estimates the solution to backward induction from training data, bridging the gap between optimal stopping theory and real-world deployment. It employs density ratio estimation and convex function learning to provide statistically consistent estimators for sufficient statistic and conditional expectation, both essential for solving backward induction; consequently,

FIRMBOUND minimizes Bayes risk to reach optimality. Additionally, we present a faster alternative using Gaussian process regression, which significantly reduces training time while retaining low deployment overhead, albeit with potential compromise in statistical consistency. Experiments across independent and identically distributed (i.i.d.), non-i.i.d., binary, multiclass, synthetic, and real-world datasets show that FIRMBOUND achieves optimalities in the sense of Bayes risk and speed-accuracy tradeoff. Furthermore, it advances the tradeoff boundary toward optimality when possible and reduces decision-time variance, ensuring reliable decision-making. Code is included in the supplementary materials.

540. Geometric Inductive Biases of Deep Networks: The Role of Data and Architecture

链接：https://iclr.cc/virtual/2025/poster/29031 abstract： In this paper, we propose the geometric invariance hypothesis (GIH), which argues that the input space curvature of a neural network remains invariant under transformation in certain architecture-dependent directions during training. We investigate a simple, non-linear binary classification problem residing on a plane in a high dimensional space and observe that—unlike MPLs—ResNets fail to generalize depending on the orientation of the plane. Motivated by this example, we define a neural network's average geometry and average geometry evolution as compact architecture-dependent summaries of the model's input-output geometry and its evolution during training. By investigating the average geometry evolution at initialization, we discover that the geometry of a neural network evolves according to the data covariance projected onto its average geometry. This means that the geometry only changes in a subset of the input space when the average geometry is low-rank, such as in ResNets. This causes an architecture-dependent invariance property in the input space curvature, which we dub GIH. Finally, we present extensive experimental results to observe the consequences of GIH and how it relates to generalization in neural networks.

541. MIND over Body: Adaptive Thinking using Dynamic Computation

链接：https://iclr.cc/virtual/2025/poster/30390 abstract： While the human brain efficiently handles various computations with a limited number of neurons, traditional deep learning networks require a significant increase in parameters to improve performance. Yet, these parameters are used inefficiently as the networks employ the same amount of computation for inputs of the same size, regardless of the input's complexity. We address this inefficiency by introducing self-introspection capabilities to the network, enabling it to adjust the number of used parameters based on the internal representation of the task and adapt the computation time based on the task complexity. This enables the network to adaptively reuse parameters across tasks, dynamically adjusting the computational effort to match the complexity of the input. We demonstrate the effectiveness of this method on language modeling and computer vision tasks. Notably, our model achieves 96.62\% accuracy on ImageNet with just a three-layer network, surpassing much larger ResNet-50 and EfficientNet. When applied to a transformer architecture, the approach achieves 95.8\%/88.7\% F1 scores on the SQuAD v1.1/v2.0 datasets at negligible parameter cost. These results showcase the potential for dynamic and reflective computation, contributing to the creation of intelligent systems that efficiently manage resources based on input data complexity.

542. Interleaved Scene Graphs for Interleaved Text-and-Image Generation Assessment

链接：https://iclr.cc/virtual/2025/poster/28211 abstract： Many real-world user queries (e.g. "How do to make egg fried rice?") could benefit from systems capable of generating responses with both textual steps with accompanying images, similar to a cookbook.Models designed to generate interleaved text and images face challenges in ensuring consistency within and across these modalities.To address these challenges, we present ISG, a comprehensive evaluation framework for interleaved text-and-image generation. ISG leverages a scene graph structure to capture relationships between text and image blocks, evaluating responses on four levels of granularity: holistic, structural, block-level, and image-specific. This multi-tiered evaluation allows for a nuanced assessment of consistency, coherence, and accuracy, and provides interpretable question-answer feedback.In conjunction with ISG, we introduce a benchmark, ISG-Bench, encompassing 1,150 samples across 8 categories and 21 subcategories. This benchmark dataset includes complex language-vision dependencies and golden answers to evaluate models effectively on vision-centric tasks such as style transfer, a challenging area for current models. Using ISG-Bench, we demonstrate that recent unified vision-language models perform poorly on generating interleaved content. While compositional approaches that combine separate language and image models show a 111% improvement over unified models at the holistic level, their performance remains suboptimal at both block and image levels.To facilitate future work, we develop ISG-Agent, a baseline agent employing a "plan-execute-refine" pipeline to invoke tools, achieving a 122% performance improvement.

543. Random-Set Neural Networks

链接：https://iclr.cc/virtual/2025/poster/28286 abstract： Machine learning is increasingly deployed in safety-critical domains where erroneous predictions may lead to potentially catastrophic consequences, highlighting the need for learning systems to be aware of how confident they are in their own predictions: in other words, 'to know when they do not know'. In this paper, we propose a novel Random-Set Neural Network (RS-NN) approach to classification which predicts belief functions (rather than classical probability vectors) over the class list using the mathematics of random sets, i.e., distributions over the collection of sets of classes. RS-NN encodes the 'epistemic' uncertainty induced by training sets that are insufficiently representative or limited in size via the size of the convex set of probability vectors associated with a predicted belief function. Our approach outperforms state-of-the-art Bayesian and Ensemble methods in terms of accuracy, uncertainty estimation and out-of-distribution

(OoD) detection on multiple benchmarks (CIFAR-10 vs SVHN/Intel-Image, MNIST vs FMNIST/KMNIST, ImageNet vs ImageNet-O). RS-NN also scales up effectively to large-scale architectures (e.g. WideResNet-28-10, VGG16, Inception V3, EfficientNetB2 and ViT-Base-16),exhibits remarkable robustness to adversarial attacks and can provide statistical guarantees in a conformal learning setting.

544. Active Learning for Continual Learning: Keeping the Past Alive in the Present

链接：https://iclr.cc/virtual/2025/poster/28450 abstract： Continual learning (CL) enables deep neural networks to adapt to ever-changing data distributions. In practice, there may be scenarios where annotation is costly, leading to active continual learning (ACL), which performs active learning (AL) for the CL scenarios when reducing the labeling cost by selecting the most informative subset is preferable. However, conventional AL strategies are not suitable for ACL, as they focus solely on learning the new knowledge, leading to catastrophic forgetting of previously learned tasks. Therefore, ACL requires a new AL strategy that can balance the prevention of catastrophic forgetting and the ability to quickly learn new tasks. In this paper, we propose AccuACL, Accumulated informativeness-based Active Continual Learning, by the novel use of the Fisher information matrix as a criterion for sample selection, derived from a theoretical analysis of the Fisher-optimality preservation properties within the framework of ACL, while also addressing the scalability issue of Fisher information-based AL. Extensive experiments demonstrate that AccuACL significantly outperforms AL baselines across various CL algorithms, increasing the average accuracy and forgetting by 23.8% and 17.0%, respectively, on average.

545. DUALFormer: Dual Graph Transformer

链接：https://iclr.cc/virtual/2025/poster/30986 abstract： Graph Transformers (GTs), adept at capturing the locality and globality of graphs, have shown promising potential in node classification tasks. Most state-of-the-art GTs succeed through integrating local Graph Neural Networks (GNNs) with their global Self-Attention (SA) modules to enhance structural awareness. Nonetheless, this architecture faces limitations arising from scalability challenges and the trade-off between capturing local and global information. On the one hand, the quadratic complexity associated with the SA modules poses a significant challenge for many GTs, particularly when scaling them to large-scale graphs. Numerous GTs necessitated a compromise, relinquishing certain aspects of their expressivity to garner computational efficiency. On the other hand, GTs face challenges in maintaining detailed local structural information while capturing long-range dependencies. As a result, they typically require significant computational costs to balance the local and global expressivity. To address these limitations, this paper introduces a novel GT architecture, dubbed DUALFormer, featuring a dual-dimensional design of its GNN and SA modules. Leveraging approximation theory from Linearized Transformers and treating the query as the surrogate representation of node features, DUALFormer \emph{efficiently} performs the computationally intensive global SA module on feature dimensions. Furthermore, by such a separation of local and global modules into dual dimensions, DUALFormer achieves a natural balance between local and global expressivity. In theory, DUALFormer can reduce intra-class variance, thereby enhancing the discriminability of node representations. Extensive experiments on eleven real-world datasets demonstrate its effectiveness and efficiency over existing state-of-the-art GTs.

546. Steering Large Language Models between Code Execution and Textual Reasoning

链接：https://iclr.cc/virtual/2025/poster/30940 abstract： While a lot of recent research focuses on enhancing the textual reasoning capabilities of Large Language Models (LLMs) by optimizing the multi-agent framework or reasoning chains, several benchmark tasks can be solved with 100\% success through direct coding, which is more scalable and avoids the computational overhead associated with textual iterating and searching. Textual reasoning has inherent limitations in solving tasks with challenges in math, logics, optimization, and searching, which is unlikely to be solved by simply scaling up the model and data size. The recently released OpenAI GPT Code Interpreter and multi-agent frameworks such as AutoGen have demonstrated remarkable proficiency of integrating code generation and execution to solve complex tasks using LLMs. However, based on our experiments on 7 existing popular methods for steering code/text generation in both single- and multi-turn settings with 14 tasks and 6 types of LLMs (including the new O1-preview), currently there is no optimal method to correctly steer LLMs to write code when needed. We discover some interesting patterns on when models use code vs. textual reasoning with the evolution to task complexity and model sizes, which even result in an astonishingly inverse scaling behavior. We also discover that results from LLM written code are not always better than using textual reasoning, even if the task could be solved through code. To mitigate the above issues, we propose three methods to better steer LLM code/text generation and achieve a notable improvement. The costs of token lengths and runtime are thoroughly discussed for all the methods. We believe the problem of steering LLM code/text generation is critical for future research and has much space for further improvement. Project Page, Datasets, and Codes are available at https://yongchao98.github.io/CodeSteer/.

547. Arithmetic Transformers Can Length-Generalize in Both Operand Length and Count

链接：https://iclr.cc/virtual/2025/poster/28933 abstract： Transformers often struggle with length generalization, meaning they fail to generalize to sequences longer than those encountered during training. While arithmetic tasks are commonly used to study

length generalization, certain tasks are considered notoriously difficult, e.g., multi-operand addition (requiring generalization over both the number of operands and their lengths) and multiplication (requiring generalization over both operand lengths). In this work, we achieve approximately 2–3× length generalization on both tasks, which is the first such achievement in arithmetic Transformers. We design task-specific scratchpads enabling the model to focus on a fixed number of tokens per each next-token prediction step, and apply multi-level versions of Position Coupling (Cho et al., 2024; McLeish et al., 2024) to let Transformers know the right position to attend to. On the theory side, we prove that a 1-layer Transformer using our method can solve multi-operand addition, up to operand length and operand count that are exponential in embedding dimension.

548. Breaking Free from MMI: A New Frontier in Rationalization by Probing Input Utilization

链接：https://iclr.cc/virtual/2025/poster/29357 abstract： Extracting a small subset of crucial rationales from the full input is a key problem in explainability research. The most widely used fundamental criterion for rationale extraction is the maximum mutual information (MMI) criterion. In this paper, we first demonstrate that MMI suffers from diminishing marginal returns. Once part of the rationale has been identified, finding the remaining portions contributes only marginally to increasing the mutual information, making it difficult to use MMI to locate the rest. In contrast to MMI that aims to reproduce the prediction, we seek to identify the parts of the input that the network can actually utilize. This is achieved by comparing how different rationale candidates match the capability space of the weight matrix. The weight matrix of a neural network is typically low-rank, meaning that the linear combinations of its column vectors can only cover part of the directions in a high-dimensional space (high-dimension: the dimensions of an input vector). If an input is fully utilized by the network, it generally matches these directions (e.g., a portion of a hypersphere), resulting in a representation with a high norm. Conversely, if an input primarily falls outside (orthogonal to) these directions, its representation norm will approach zero, behaving like noise that the network cannot effectively utilize. Building on this, we propose using the norms of rationale candidates as an alternative objective to MMI. Through experiments on four text classification datasets and one graph classification dataset using three network architectures (GRUs, BERT, and GCN), we show that our method outperforms MMI and its improved variants in identifying better rationales. We also compare our method with a representative LLM (llama-3.1-8b-instruct) and find that our simple method gets comparable results to it and can sometimes even outperform it.

549. Convergence and Implicit Bias of Gradient Descent on Continual Linear Classification

链接：https://iclr.cc/virtual/2025/poster/30455 abstract： We study continual learning on multiple linear classification tasks by sequentially running gradient descent (GD) for a fixed budget of iterations per each given task. When all tasks are jointly linearly separable and are presented in a cyclic/random order, we show the directional convergence of the trained linear classifier to the joint (offline) max-margin solution. This is surprising because GD training on a single task is implicitly biased towards the individual max-margin solution for the task, and the direction of the joint max-margin solution can be largely different from these individual solutions. Additionally, when tasks are given in a cyclic order, we present a non-asymptotic analysis on cycle-averaged forgetting, revealing that (1) alignment between tasks is indeed closely tied to catastrophic forgetting and backward knowledge transfer and (2) the amount of forgetting vanishes to zero as the cycle repeats. Lastly, we analyze the case where the tasks are no longer jointly separable and show that the model trained in a cyclic order converges to the unique minimum of the joint loss function.

550. 3D-Properties: Identifying Challenges in DPO and Charting a Path Forward

链接：https://iclr.cc/virtual/2025/poster/30704 abstract： Aligning large language models (LLMs) with human preferences has gained significant attention, with Proximal Policy Optimization (PPO) as a standard yet computationally expensive method and Direct Preference Optimization (DPO) as a more efficient alternative. While DPO offers simplicity, it remains underutilized in state-of-the-art LLMs, suggesting potential limitations. In this work, we revisit DPO, analyzing its theoretical foundations and empirical performance to bridge this gap. We identify three key properties—termed \textbf{3D}-properties—that emerge from DPO’s learning process: \textbf{D}rastic drop in rejected response likelihood, \textbf{D}egradation into response suppression, and \textbf{D}ispersion effect on unseen responses. We show that these issues arise from DPO’s optimization dynamics, where the interaction between chosen and rejected response gradients leads to instability. Our findings are supported by experiments on both a controlled toy model and real-world LLM tasks, including mathematical problem-solving and instruction following. To address these challenges, we propose simple regularization techniques that improve training stability and performance. Additionally, we examine how preference data distribution impacts DPO’s effectiveness, offering insights into how alignment models handle out-of-domain (OOD) data. Our work connects these observations to broader research and provides a theoretical explanation for DPO’s limitations. We hope these insights will guide future advancements in reward-model-free preference learning, bringing it closer to reward-model-based approaches.

551. Leave-One-Out Stable Conformal Prediction

链接：https://iclr.cc/virtual/2025/poster/30540 abstract： Conformal prediction (CP) is an important tool for distribution-free predictive uncertainty quantification.Yet, a major challenge is to balance computational efficiency and prediction accuracy,

particularly for multiple predictions.We propose Leave-One-Out Stable Conformal Prediction (LOO-StabCP), a novel method to speed up full conformal using algorithmic stability without sample splitting.By leveraging leave-one-out stability, our method is much faster in handling a large number of prediction requests compared to existing method RO-StabCP based on replace-one stability.We derived stability bounds for several popular machine learning tools: regularized loss minimization (RLM) and stochastic gradient descent (SGD), as well as kernel method, neural networks and bagging.Our method is theoretically justified and demonstrates superior numerical performance on synthetic and real-world data.We applied our method to a screening problem, where its effective exploitation of training data led to improved test power compared to state-of-the-art method based on split conformal.

552. DeepGate4: Efficient and Effective Representation Learning for Circuit Design at Scale

链接：https://iclr.cc/virtual/2025/poster/29131 abstract： Circuit representation learning has become pivotal in electronic design automation, enabling critical tasks such as testability analysis, logic reasoning, power estimation, and SAT solving. However, existing models face significant challenges in scaling to large circuits due to limitations like over-squashing in graph neural networks and the quadratic complexity of transformer-based models. To address these issues, we introduce \textbf{DeepGate4}, a scalable and efficient graph transformer specifically designed for large-scale circuits. DeepGate4 incorporates several key innovations: (1) an update strategy tailored for circuit graphs, which reduce memory complexity to sub-linear and is adaptable to any graph transformer; (2) a GAT-based sparse transformer with global and local structural encodings for AIGs; and (3) an inference acceleration CUDA kernel that fully exploit the unique sparsity patterns of AIGs. Our extensive experiments on the ITC99 and EPFL benchmarks show that DeepGate4 significantly surpasses state-of-the-art methods, achieving 15.5\% and 31.1\% performance improvements over the next-best models. Furthermore, the Fused-DeepGate4 variant reduces runtime by 35.1\% and memory usage by 46.8\%, making it highly efficient for large-scale circuit analysis. These results demonstrate the potential of DeepGate4 to handle complex EDA tasks while offering superior scalability and efficiency.

553. GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks

链接：https://iclr.cc/virtual/2025/poster/30925 abstract： Understanding complex three-dimensional (3D) structures of graphs is essential for accurately modeling various properties, yet many existing approaches struggle with fully capturing the intricate spatial relationships and symmetries inherent in such systems, especially in large-scale, dynamic molecular datasets. These methods often must balance trade-offs between expressiveness and computational efficiency, limiting their scalability. To address this gap, we propose a novel Geometric Tensor Network (GotenNet) that effectively models the geometric intricacies of 3D graphs while ensuring strict equivariance under the Euclidean group E(3). Our approach directly tackles the expressiveness-efficiency trade-off by leveraging effective geometric tensor representations without relying on irreducible representations or Clebsch-Gordan transforms, thereby reducing computational overhead. We introduce a unified structural embedding, incorporating geometry-aware tensor attention and hierarchical tensor refinement that iteratively updates edge representations through inner product operations on high-degree steerable features, allowing for flexible and efficient representations for various tasks. We evaluated models on QM9, rMD17, MD22, and Molecule3D datasets, where the proposed model consistently outperforms state-of-the-art methods in both scalar and high-degree property predictions, demonstrating exceptional robustness across diverse datasets, and establishes GotenNet as a versatile and scalable framework for 3D equivariant Graph Neural Networks.

554. Visually Guided Decoding: Gradient-Free Hard Prompt Inversion with Language Models

链接：https://iclr.cc/virtual/2025/poster/28466 abstract： Text-to-image generative models like DALL-E and Stable Diffusion have revolutionized visual content creation across various applications, including advertising, personalized media, and design prototyping.However, crafting effective textual prompts to guide these models remains challenging, often requiring extensive trial and error. Existing prompt inversion approaches, such as soft and hard prompt techniques, are not so effective due to the limited interpretability and incoherent prompt generation. To address these issues, we propose Visually Guided Decoding (VGD), a gradient-free approach that leverages large language models (LLMs) and CLIP-based guidance to generate coherent and semantically aligned prompts. In essence, VGD utilizes the robust text generation capabilities of LLMs to produce human-readable prompts. Further, by employing CLIP scores to ensure alignment with user-specified visual concepts, VGD enhances the interpretability, generalization, and flexibility of prompt generation without the need for additional training. Our experiments demonstrate that VGD outperforms existing prompt inversion techniques in generating understandable and contextually relevant prompts, facilitating more intuitive and controllable interactions with text-to-image models.

555. Tight Time Complexities in Parallel Stochastic Optimization with Arbitrary Computation Dynamics

链接：https://iclr.cc/virtual/2025/poster/29045 abstract： In distributed stochastic optimization, where parallel and asynchronous methods are employed, we establish optimal time complexities under virtually any computation behavior of workers/devices/CPUs/GPUs, capturing potential disconnections due to hardware and network delays, time-varying computation powers, and any possible fluctuations and trends of computation speeds. These real-world scenarios are

formalized by our new universal computation model. Leveraging this model and new proof techniques, we discover tight lower bounds that apply to virtually all synchronous and asynchronous methods, including Minibatch SGD, Asynchronous SGD (Recht et al., 2011), and Picky SGD (Cohen et al., 2021). We show that these lower bounds, up to constant factors, are matched by the optimal Rennala SGD and Malenia SGD methods (Tyurin & Richtárik, 2023).

556. Swift4D: Adaptive divide-and-conquer Gaussian Splatting for compact and efficient reconstruction of dynamic scene

链接：https://iclr.cc/virtual/2025/poster/29075 abstract： Novel view synthesis has long been a practical but challenging task, although the introduction of numerous methods to solve this problem, even combining advanced representations like 3D Gaussian Splatting, they still struggle to recover high-quality results and often consume too much storage memory and training time. In this paper we propose Swift4D, a divide-and-conquer 3D Gaussian Splatting method that can handle static and dynamic primitives separately, achieving a good trade-off between rendering quality and efficiency, motivated by the fact that most of the scene is the static primitive and does not require additional dynamic properties. Concretely, we focus on modeling dynamic transformations only for the dynamic primitives which benefits both efficiency and quality. We first employ a learnable decomposition strategy to separate the primitives, which relies on an additional parameter to classify primitives as static or dynamic. For the dynamic primitives, we employ a compact multi-resolution 4D Hash mapper to transform these primitives from canonical space into deformation space at each timestamp, and then mix the static and dynamic primitives to produce the final output. This divide-and-conquer method facilitates efficient training and reduces storage redundancy. Our method not only achieves state-of-the-art rendering quality while being 20× faster in training than previous SOTA methods with a minimum storage requirement of only 30MB on real-world datasets.

557. PortLLM: Personalizing Evolving Large Language Models with Training-Free and Portable Model Patches

链接：https://iclr.cc/virtual/2025/poster/28792 abstract： As large language models (LLMs) increasingly shape the AI landscape, fine-tuning pretrained models has become more popular than in the pre-LLM era for achieving optimal performance in domain-specific tasks. However, pretrained LLMs such as ChatGPT are periodically evolved (i.e., model parameters are frequently updated), making it challenging for downstream users with limited resources to keep up with fine-tuning the newest LLMs for their domain application. Even though fine-tuning costs have nowadays been reduced thanks to the innovations of parameter-efficient fine-tuning such as LoRA, not all downstream users have adequate computing for frequent personalization. Moreover, access to fine-tuning datasets, particularly in sensitive domains such as healthcare, could be time-restrictive, making it crucial to retain the knowledge encoded in earlier fine-tuned rounds for future adaptation. In this paper, we present PORTLLM, a training-free framework that (i) creates an initial lightweight model update patch to capture domain-specific knowledge, and (ii) allows a subsequent seamless plugging for the continual personalization of evolved LLM at minimal cost. Our extensive experiments cover seven representative datasets, from easier question-answering tasks {BoolQ, SST2} to harder reasoning tasks {WinoGrande, GSM8K}, and models including {Mistral-7B,Llama2, Llama3.1, and Gemma2}, validating the portability of our designed model patches and showcasing the effectiveness of our proposed framework. For instance, PORTLLM achieves comparable performance to LoRA fine-tuning with reductions of up to 12.2× in GPU memory usage. Finally, we provide theoretical justifications to understand the portability of our model update patches, which offers new insights into the theoretical dimension of LLMs' personalization.

558. Provable Benefit of Annealed Langevin Monte Carlo for Non-log-concave Sampling

链接：https://iclr.cc/virtual/2025/poster/29777 abstract： We consider the outstanding problem of sampling from an unnormalized density that may be non-log-concave and multimodal. To enhance the performance of simple Markov chain Monte Carlo (MCMC) methods, techniques of annealing type have been widely used. However, quantitative theoretical guarantees of these techniques are under-explored. This study takes a first step toward providing a non-asymptotic analysis of annealed MCMC. Specifically, we establish, for the first time, an oracle complexity of $\widetilde{O}\left(\frac{d\beta^2{\cal A}^2}{\varepsilon^6}\right)$ for the simple annealed Langevin Monte Carlo algorithm to achieve ε^2 accuracy in Kullback-Leibler divergence to the target distribution $\pi\propto{\rm e}^{-V}$ on $\mathbb{R}^d$ with β-smooth potential V. Here, ${\cal A}$ represents the action of a curve of probability measures interpolating the target distribution π and a readily sampleable distribution.

559. SeedLM: Compressing LLM Weights into Seeds of Pseudo-Random Generators

链接：https://iclr.cc/virtual/2025/poster/28000 abstract： Large Language Models (LLMs) have transformed natural language processing, but face significant challenges in widespread deployment due to their high runtime cost. In this paper, we introduce SeedLM, a novel post-training compression method that uses seeds of a pseudo-random generator to encode and compress model weights. Specifically, for each block of weights, we find a seed that is fed into a Linear Feedback Shift Register (LFSR) during inference to efficiently generate a random matrix. This matrix is then linearly combined with compressed coefficients to reconstruct the weight block. SeedLM reduces memory access and leverages idle compute cycles during inference, effectively

speeding up memory-bound tasks by trading compute for fewer memory accesses. Unlike state-of-the-art methods that rely on calibration data, our approach is data-free and generalizes well across diverse tasks. Our experiments with Llama3 70B, which is particularly challenging, show zero-shot accuracy retention at 4- and 3-bit compression to be on par with or better than state-of-the-art methods, while maintaining performance comparable to FP16 baselines. Additionally, FPGA-based tests demonstrate that 4-bit SeedLM, as model size increases, approaches a 4x speed-up over an FP16 Llama 2/3 baseline.

560. Knowledge Distillation with Multi-granularity Mixture of Priors for Image Super-Resolution

链接：https://iclr.cc/virtual/2025/poster/29042 abstract： Knowledge distillation (KD) is a promising yet challenging model compression approach that transmits rich learning representations from robust but resource-demanding teacher models to efficient student models. Previous methods for image super-resolution (SR) are often tailored to specific teacher-student architectures, limiting their potential for improvement and hindering broader applications. This work presents a novel KD framework for SR models, the multi-granularity Mixture of Priors Knowledge Distillation (MiPKD), which can be universally applied to a wide range of architectures at both feature and block levels. The teacher’s knowledge is effectively integrated with the student's feature via the Feature Prior Mixer, and the reconstructed feature propagates dynamically in the training phase with the Block Prior Mixer. Extensive experiments illustrate the significance of the proposed MiPKD technique.

561. Control-oriented Clustering of Visual Latent Representation

链接：https://iclr.cc/virtual/2025/poster/28302 abstract： We initiate a study of the geometry of the visual representation space ---the information channel from the vision encoder to the action decoder--- in an image-based control pipeline learned from behavior cloning. Inspired by the phenomenon of neural collapse (NC) in image classification, we empirically demonstrate the prevalent emergence of a similar law of clustering in the visual representation space. Specifically, - In discrete image-based control (e.g., Lunar Lander), the visual representations cluster according to the natural discrete action labels;- In continuous image-based control (e.g., Planar Pushing and Block Stacking), the clustering emerges according to ``control-oriented'' classes that are based on (a) the relative pose between the object and the target in the input or (b) the relative pose of the object induced by expert actions in the output. Each of the classes corresponds to one relative pose orthant (REPO).Beyond empirical observation, we show such a law of clustering can be leveraged as an algorithmic tool to improve test-time performance when training a policy with limited expert demonstrations. Particularly, we pretrain the vision encoder using NC as a regularization to encourage control-oriented clustering of the visual features. Surprisingly, such an NC-pretrained vision encoder, when finetuned end-to-end with the action decoder, boosts the test-time performance by 10% to 35%. Real-world vision-based planar pushing experiments confirmed the surprising advantage of control-oriented visual representation pretraining.

562. Implicit Search via Discrete Diffusion: A Study on Chess

链接：https://iclr.cc/virtual/2025/poster/30648 abstract： In the post-AlphaGo era, there has been a renewed interest in search techniques such as Monte Carlo Tree Search (MCTS), particularly in their application to Large Language Models (LLMs).This renewed attention is driven by the recognition that current next-token prediction models often lack the ability for long-term planning. Is it possible to instill search-like abilities within the models to enhance their planning abilities without relying on explicit search? We propose DiffuSearch , a model that does \textit{implicit search} by looking into the future world via discrete diffusion modeling. We instantiate DiffuSearch on a classical board game, Chess, where explicit search is known to be essential. Through extensive controlled experiments, we show DiffuSearch outperforms both the searchless and explicit search-enhanced policies. Specifically, DiffuSearch outperforms the one-step policy by 19.2\% and the MCTS-enhanced policy by 14\% on action accuracy. Furthermore, DiffuSearch demonstrates a notable 30\% enhancement in puzzle-solving abilities compared to explicit search-based policies, along with a significant 540 Elo increase in game-playing strength assessment. These results indicate that implicit search via discrete diffusion is a viable alternative to explicit search over a one-step policy. All codes are publicly available at \href{https://github.com/HKUNLP/DiffuSearch}{https://github.com/HKUNLP/DiffuSearch}.

563. Ensembles of Low-Rank Expert Adapters

链接：https://iclr.cc/virtual/2025/poster/28544 abstract： The training and fine-tuning of large language models (LLMs) often involve diverse textual data from multiple sources, which poses challenges due to conflicting gradient directions, hindering optimization and specialization. These challenges can undermine model generalization across tasks, resulting in reduced downstream performance. Recent research suggests that fine-tuning LLMs on carefully selected, task-specific subsets of data can match or even surpass the performance of using the entire dataset. Building on these insights, we propose the Ensembles of Low-Rank Expert Adapters (ELREA) framework to improve the model's capability to handle diverse tasks. ELREA clusters the training instructions based on their gradient directions, representing different areas of expertise and thereby reducing conflicts during optimization. Expert adapters are then trained on these clusters, utilizing the low-rank adaptation (LoRA) technique to ensure training efficiency and model scalability. During inference, ELREA combines predictions from the most relevant expert adapters based on the input data's gradient similarity to the training clusters, ensuring optimal adapter selection for each task. Experiments show that our method outperforms baseline LoRA adapters trained on the full dataset and other ensemble approaches with similar training and inference complexity across a range of domain-specific tasks.

564. Interactive Speculative Planning: Enhance Agent Efficiency through Co-

design of System and User Interface

链接：https://iclr.cc/virtual/2025/poster/30539 abstract： Agents, as user-centric tools, are increasingly deployed for human task delegation, assisting with a broad spectrum of requests by generating thoughts, engaging with user proxies, and producing action plans. However, agents based on large language models often face substantial planning latency due to two primary factors: the efficiency limitations of the underlying LLMs due to their large size and high demand, and the structural complexity of the agents due to the extensive generation of intermediate steps to produce the final output. Given that inefficiency in service provision can undermine the value of automation for users, this paper presents a human-centered efficient agent planning method – Interactive Speculative Planning – aiming at enhancing the efficiency of agent planning through both system design and user interaction. Our approach advocates for the co-design of the agent system and user interface, underscoring the importance of an agent system that can fluidly manage user interactions and interruptions. By integrating human interruptions as a fundamental component of the system, we not only make it more user-centric but also expedite the entire process by leveraging human-in-the-loop interactions to provide accurate intermediate steps.

565. OS-ATLAS: Foundation Action Model for Generalist GUI Agents

链接：https://iclr.cc/virtual/2025/poster/28423 abstract： Existing efforts in building GUI agents heavily rely on the availability of robust commercial Vision-Language Models (VLMs) such as GPT-4o and GeminiProVision. Practitioners are often reluctant to use open-source VLMs due to their significant performance lag compared to their closed-source counterparts, particularly in GUI grounding and Out-Of-Distribution (OOD) scenarios. To facilitate future research in this area, we developed OS-Atlas—a foundational GUI action model that excels at GUI grounding and OOD agentic tasks through innovations in both data and modeling.We have invested significant engineering effort in developing an open-source toolkit for synthesizing GUI grounding data across multiple platforms, including Windows, Linux, MacOS, Android, and the web. Leveraging this toolkit, we are releasing the largest open-source cross-platform GUI grounding corpus to date, which contains over 13 million GUI elements. This dataset, combined with innovations in model training, provides a solid foundation for OS-Atlas to understand GUI screenshots and generalize to unseen interfaces.Through extensive evaluation across six benchmarks spanning three different platforms (mobile, desktop, and web), OS-Atlas demonstrates significant performance improvements over previous state-of-the-art models. Our evaluation also uncovers valuable insights into continuously improving and scaling the agentic capabilities of open-source VLMs.

566. MMSearch: Unveiling the Potential of Large Models as Multi-modal Search Engines

链接：https://iclr.cc/virtual/2025/poster/30134 abstract： The advent of Large Language Models (LLMs) has paved the way for AI search engines, e.g., SearchGPT, showcasing a new paradigm in human-internet interaction. However, most current AI search engines are limited to text-only settings, neglecting the multimodal user queries and the text-image interleaved nature of website information. Recently, Large Multimodal Models (LMMs) have made impressive strides. Yet, whether they can function as AI search engines remains under-explored, leaving the potential of LMMs in multimodal search an open question. To this end, we first design a delicate pipeline, MMSearch-Engine, to empower any LMMs with multimodal search capabilities. On top of this, we introduce MMSearch, a comprehensive evaluation benchmark to assess the multimodal search performance of LMMs. The curated dataset contains 300 manually collected instances spanning 14 subfields, which involves no overlap with the current LMMs' training data, ensuring the correct answer can only be obtained within searching. By using MMSearch-Engine, the LMMs are evaluated by performing three individual tasks (requery, rerank, and summarization), and one challenging end-to-end task with a complete searching process. We conduct extensive experiments on closed-source and open-source LMMs. Among all tested models, GPT-4o with MMSearch-Engine achieves the best results, which surpasses the commercial product, Perplexity Pro, in the end-to-end task, demonstrating the effectiveness of our proposed pipeline. We further present error analysis to unveil current LMMs still struggle to fully grasp the multimodal search tasks, and conduct ablation study to indicate the potential of scaling test-time computation for AI search engine. We hope MMSearch may provide unique insights to guide the future development of multimodal AI search engine.

567. Token Statistics Transformer: Linear-Time Attention via Variational Rate Reduction

链接：https://iclr.cc/virtual/2025/poster/28513 abstract： The attention operator is arguably the key distinguishing factor of transformer architectures, which have demonstrated state-of-the-art performance on a variety of tasks. However, transformer attention operators often impose a significant computational burden, with the computational complexity scaling quadratically with the number of tokens. In this work, we propose a novel transformer attention operator whose computational complexity scales linearly with the number of tokens. We derive our network architecture by extending prior work which has shown that a transformer style architecture naturally arises by "white-box" architecture design, where each layer of the network is designed to implement an incremental optimization step of a maximal coding rate reduction objective (MCR^2). Specifically, we derive a novel variational form of the MCR^2 objective and show that the architecture that results from unrolled gradient descent of this variational objective leads to a new attention module called Token Statistics Self-Attention ($\texttt{TSSA}$). $\texttt{TSSA}$ has $\textit{linear computational and memory complexity}$ and radically departs from the typical attention architecture that computes pairwise similarities between tokens. Experiments on vision, language, and long sequence tasks show that simply swapping $\texttt{TSSA}$ for standard self-attention, which we refer to as the Token Statistics Transformer ($\texttt{ToST}$), achieves

competitive performance with conventional transformers while being significantly more computationally efficient and interpretable. Our results also somewhat call into question the conventional wisdom that pairwise similarity style attention mechanisms are critical to the success of transformer architectures.

568. RFWave: Multi-band Rectified Flow for Audio Waveform Reconstruction

链接：https://iclr.cc/virtual/2025/poster/28820 abstract： Recent advancements in generative modeling have significantly enhanced the reconstruction of audio waveforms from various representations. While diffusion models are adept at this task, they are hindered by latency issues due to their operation at the individual sample point level and the need for numerous sampling steps. In this study, we introduce RFWave, a cutting-edge multi-band Rectified Flow approach designed to reconstruct high-fidelity audio waveforms from Mel-spectrograms or discrete acoustic tokens. RFWave uniquely generates complex spectrograms and operates at the frame level, processing all subbands simultaneously to boost efficiency. Leveraging Rectified Flow, which targets a straight transport trajectory, RFWave achieves reconstruction with just 10 sampling steps. Our empirical evaluations show that RFWave not only provides outstanding reconstruction quality but also offers vastly superior computational efficiency, enabling audio generation at speeds up to 160 times faster than real-time on a GPU. Both an online demonstration and the source code are accessible.

569. Apollo-MILP: An Alternating Prediction-Correction Neural Solving Framework for Mixed-Integer Linear Programming

链接：https://iclr.cc/virtual/2025/poster/28474 abstract： Leveraging machine learning (ML) to predict an initial solution for mixed-integer linear programming (MILP) has gained considerable popularity in recent years. These methods predict a solution and fix a subset of variables to reduce the problem dimension. Then, they solve the reduced problem to obtain the final solutions. However, directly fixing variable values can lead to low-quality solutions or even infeasible reduced problems if the predicted solution is not accurate enough. To address this challenge, we propose an Alternating prediction-correction neural solving framework (Apollo-MILP) that can identify and select accurate and reliable predicted values to fix. In each iteration, Apollo-MILP conducts a prediction step for the unfixed variables, followed by a correction step to obtain an improved solution (called reference solution) through a trust-region search. By incorporating the predicted and reference solutions, we introduce a novel Uncertainty-based Error upper BOund (UEBO) to evaluate the uncertainty of the predicted values and fix those with high confidence. A notable feature of Apollo-MILP is the superior ability for problem reduction while preserving optimality, leading to high-quality final solutions. Experiments on commonly used benchmarks demonstrate that our proposed Apollo-MILP significantly outperforms other ML-based approaches in terms of solution quality, achieving over a 50% reduction in the solution gap.

570. PerturboLLaVA: Reducing Multimodal Hallucinations with Perturbative Visual Training

链接：https://iclr.cc/virtual/2025/poster/28657 abstract： This paper aims to address the challenge of hallucinations in Multimodal Large Language Models (MLLMs) particularly for dense image captioning tasks. To tackle the challenge, we identify the current lack of a metric that finely measures the caption quality in concept level. We hereby introduce HalFscore, a novel metric built upon the language graph and is designed to evaluate both the accuracy and completeness of dense captions at agranular level. Additionally, we identify the root cause of hallucination as the model's over-reliance on its language prior. To address this, we propose PerturboLLaVA, which reduces the model's reliance on the language prior by incorporating adversarially perturbed text during training. This method enhances the model's focus on visual inputs, effectively reducing hallucinations and producing accurate, image-grounded descriptions without incurring additional computational overhead. PerturboLLaVA significantly improves the fidelity of generated captions, outperforming existing approaches in handling multimodal hallucinations and achieving improved performance across general multimodal benchmarks.

571. Robust Watermarking Using Generative Priors Against Image Editing: From Benchmarking to Advances

链接：https://iclr.cc/virtual/2025/poster/31223 abstract： Current image watermarking methods are vulnerable to advanced image editing techniques enabled by large-scale text-to-image models. These models can distort embedded watermarks during editing, posing significant challenges to copyright protection. In this work, we introduce W-Bench, the first comprehensive benchmark designed to evaluate the robustness of watermarking methods against a wide range of image editing techniques, including image regeneration, global editing, local editing, and image-to-video generation. Through extensive evaluations of eleven representative watermarking methods against prevalent editing techniques, we demonstrate that most methods fail to detect watermarks after such edits. To address this limitation, we propose VINE, a watermarking method that significantly enhances robustness against various image editing techniques while maintaining high image quality. Our approach involves two key innovations: (1) we analyze the frequency characteristics of image editing and identify that blurring distortions exhibit similar frequency properties, which allows us to use them as surrogate attacks during training to bolster watermark robustness; (2) we leverage a large-scale pretrained diffusion model SDXL-Turbo, adapting it for the watermarking task to achieve more imperceptible and robust watermark embedding. Experimental results show that our method achieves outstanding watermarking performance under various image editing techniques, outperforming existing methods in both image quality and robustness.

Code is available at https://github.com/Shilin-LU/VINE

572. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation

链接：https://iclr.cc/virtual/2025/poster/31274 abstract： VILA-U is a Unified foundation model that integrates Video, Image, Language understanding and generation. Traditional visual language models (VLMs) use separate modules for understanding and generating visual content, which can lead to misalignment and increased complexity. In contrast, VILA-U employs a single autoregressive next-token prediction framework for both tasks, eliminating the need for additional components like diffusion models. This approach not only simplifies the model but also achieves near state-of-the-art performance in visual language understanding and generation. The success of VILA-U is attributed to two main factors: the unified vision tower that aligns discrete visual tokens with textual inputs during pretraining, which enhances visual perception, and autoregressive image generation can achieve similar quality as diffusion models with high-quality dataset. This allows VILA-U to perform comparably to more complex models using a fully token-based autoregressive framework.

573. HART: Efficient Visual Generation with Hybrid Autoregressive Transformer

链接：https://iclr.cc/virtual/2025/poster/32057 abstract： We introduce Hybrid Autoregressive Transformer (HART), the first autoregressive (AR) visual generation model capable of directly generating 1024x1024 images, rivaling diffusion models in image generation quality. Existing AR models face limitations due to the poor image reconstruction quality of their discrete tokenizers and the prohibitive training costs associated with generating 1024px images. To address these challenges, we present the hybrid tokenizer, which decomposes the continuous latents from the autoencoder into two components: discrete tokens representing the big picture and continuous tokens representing the residual components that cannot be represented by the discrete tokens. The discrete component is modeled by a scalable-resolution discrete AR model, while the continuous component is learned with a lightweight residual diffusion module with only 37M parameters. Compared with the discrete-only VAR tokenizer, our hybrid approach improves reconstruction FID from 2.11 to 0.30 on MJHQ-30K, leading to a 31% generation FID improvement from 7.85 to 5.38. HART also outperforms state-of-the-art diffusion models in both FID and CLIP score, with 4.5-7.7$\times$ higher throughput and 6.9-13.4$\times$ lower MACs. Our code is open sourced at https://github.com/mit-han-lab/hart.

574. SPAM: Spike-Aware Adam with Momentum Reset for Stable LLM Training

链接：https://iclr.cc/virtual/2025/poster/30015 abstract： Large Language Models (LLMs) have demonstrated exceptional performance across diverse tasks, yet their training remains highly resource intensive and susceptible to critical challenges such as training instability. A predominant source of this instability stems from gradient and loss spikes, which disrupt the learning process, often leading to costly interventions like checkpoint recovery and experiment restarts, further amplifying inefficiencies. This paper presents a comprehensive investigation into gradient spikes observed during LLM training, revealing their prevalence across multiple architectures and datasets. Our analysis shows that these spikes can be up to 1000× larger than typical gradients, substantially deteriorating model performance. To address this issue, we propose Spike-Aware Adam with Momentum Reset (SPAM), a novel optimizer designed to counteract gradient spikes through momentum reset and spike-aware gradient clipping. Extensive experiments, including both pre-training and fine-tuning, demonstrate that SPAM consistently surpasses Adam and its variants across a range of model scales. Additionally, SPAM facilitates memory-efficient training by enabling sparse momentum, where only a subset of momentum terms are maintained and updated. When operating under memory constraints, SPAM outperforms state-of-the-art memory-efficient optimizers such as GaLore and Adam-Mini. Our work underscores the importanceof mitigating gradient spikes in LLM training and introduces an effective optimization strategy that enhances both training stability and resource efficiency at scale. Code is submitted.

575. Classic but Everlasting: Traditional Gradient-Based Algorithms Converge Fast Even in Time-Varying Multi-Player Games

链接：https://iclr.cc/virtual/2025/poster/28077 abstract： Last-iterate convergence behaviours of well-known algorithms are intensively investigated in various games, such as two-player bilinear zero-sum games.However, most known last-iterate convergence properties rely on strict settings where the underlying games must have time-invariant payoffs.Besides, the limited known attempts on the games with time-varying payoffs are in two-player bilinear time-varying zero-sum games and strictly monotone games. By contrast, in other time-varying games, the last-iterate behaviours of two classic algorithms, i.e., extra gradient (EG) and optimistic gradient (OG) algorithms, still lack research, especially the convergence rates in multi-player games.In this paper, we investigate the last-iterate behaviours of EG and OG algorithms for convergent perturbed games, which extend upon the usual model of time-invariant games and incorporate external factors, such as vanishing noises.Using the recently proposed notion of the tangent residual (or its modifications) as the potential function of games and the measure of proximity to the Nash equilibrium, we prove that the last-iterate convergence rates of EG and OG algorithms for perturbed games on bounded convex closed sets are $O({1}/{\sqrt{T}})$ if such games converge to monotone games at rates fast enough and that such a result holds true for certain unconstrained perturbed games. With this result, we address an open questionasking

for the last-iterate convergence rate of EG and OG algorithms in constrained and time-varying settings. The above convergence rates are similar to known tight results on corresponding time-invariant games.

576. I Can Hear You: Selective Robust Training for Deepfake Audio Detection

链接：https://iclr.cc/virtual/2025/poster/32111 abstract： Recent advances in AI-generated voices have intensified the challenge of detecting deepfake audio, posing risks for scams and the spread of disinformation. To tackle this issue, we establish the largest public voice dataset to date, named DeepFakeVox-HQ, comprising 1.3 million samples, including 270,000 high-quality deepfake samples from 14 diverse sources. Despite previously reported high accuracy, existing deepfake voice detectors struggle with our diversely collected dataset, and their detection success rates drop even further under realistic corruptions and adversarial attacks. We conduct a holistic investigation into factors that enhance model robustness and show that incorporating a diversified set of voice augmentations is beneficial. Moreover, we find that the best detection models often rely on high-frequency features, which are imperceptible to humans and can be easily manipulated by an attacker. To address this, we propose the F-SAT: Frequency-Selective Adversarial Training method focusing on high-frequency components. Empirical results demonstrate that using our training dataset boosts baseline model performance (without robust training) by 33%, and our robust training further improves accuracy by 7.7% on clean samples and by 29.3% on corrupted and attacked samples, over the state-of-the-art RawNet3 model.

577. (Mis)Fitting Scaling Laws: A Survey of Scaling Law Fitting Techniques in Deep Learning

链接：https://iclr.cc/virtual/2025/poster/27795 abstract： Modern foundation models rely heavily on using scaling laws to guide crucial training decisions. Researchers often extrapolate the optimal architecture and hyper parameters settings from smaller training runs by describing the relationship between, loss, or task performance, and scale. All components of this process vary, from the specific equation being fit, to the training setup, to the optimization method. Each of these factors may affect the fitted law, and therefore, the conclusions of a given study. We discuss discrepancies in the conclusions that several prior works reach, on questions such as the optimal token to parameter ratio. We augment this discussion with our own analysis of the critical impact that changes in specific details may effect in a scaling study, and the resulting altered conclusions. Additionally, we survey over 50 papers that study scaling trends: while 45 of these papers quantify these trends using a power law, most under-report crucial details needed to reproduce their findings. To mitigate this, we we propose a checklist for authors to consider while contributing to scaling law research.

578. Generating CAD Code with Vision-Language Models for 3D Designs

链接：https://iclr.cc/virtual/2025/poster/30576 abstract： Generative AI has transformed the fields of Design and Manufacturing by providingefficient and automated methods for generating and modifying 3D objects. Oneapproach involves using Large Language Models (LLMs) to generate Computer-Aided Design (CAD) scripting code, which can then be executed to render a 3Dobject; however, the resulting 3D object may not meet the specified requirements.Testing the correctness of CAD generated code is challenging due to the complexityand structure of 3D objects (e.g., shapes, surfaces, and dimensions) that are notfeasible in code. In this paper, we introduce CADCodeVerify, a novel approach toiteratively verify and improve 3D objects generated from CAD code. Our approachworks by producing ameliorative feedback by prompting a Vision-Language Model(VLM) to generate and answer a set of validation questions to verify the generatedobject and prompt the VLM to correct deviations. To evaluate CADCodeVerify, weintroduce, CADPrompt, the first benchmark for CAD code generation, consisting of200 natural language prompts paired with expert-annotated scripting code for 3Dobjects to benchmark progress. Our findings show that CADCodeVerify improvesVLM performance by providing visual feedback, enhancing the structure of the 3Dobjects, and increasing the success rate of the compiled program. When applied toGPT-4, CADCodeVerify achieved a 7.30% reduction in Point Cloud distance and a5.0% improvement in success rate compared to prior work.

579. An Optimal Discriminator Weighted Imitation Perspective for Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/30701 abstract： We introduce Iterative Dual Reinforcement Learning (IDRL), a new method that takes an optimal discriminator-weighted imitation view of solving RL. Our method is motivated by a simple experiment in which we find training a discriminator using the offline dataset plus an additional expert dataset and then performing discriminator-weighted behavior cloning gives strong results on various types of datasets. That optimal discriminator weight is quite similar to the learned visitation distribution ratio in Dual-RL, however, we find that current Dual-RL methods do not correctly estimate that ratio. In IDRL, we propose a correction method to iteratively approach the optimal visitation distribution ratio in the offline dataset given no addtional expert dataset. During each iteration, IDRL removes zero-weight suboptimal transitions using the learned ratio from the previous iteration and runs Dual-RL on the remaining subdataset. This can be seen as replacing the behavior visitation distribution with the optimized visitation distribution from the previous iteration, which theoretically gives a curriculum of improved visitation distribution ratios that are closer to the optimal discriminator weight. We verify the effectiveness of IDRL on various kinds of offline datasets, including D4RL datasets and more realistic corrupted demonstrations. IDRL beats strong Primal-RL and Dual-RL baselines in terms of both performance and stability, on all datasets.

580. A-Bench: Are LMMs Masters at Evaluating AI-generated Images?

链接：https://iclr.cc/virtual/2025/poster/30994 abstract： How to accurately and efficiently assess AI-generated images (AIGIs) remains a critical challenge for generative models. Given the high costs and extensive time commitments required for user studies, many researchers have turned towards employing large multi-modal models (LMMs) as AIGI evaluators, the precision and validity of which are still questionable. Furthermore, traditional benchmarks often utilize mostly natural-captured content rather than AIGIs to test the abilities of LMMs, leading to a noticeable gap for AIGIs. Therefore, we introduce A-Bench in this paper, a benchmark designed to diagnose whether LMMs are masters at evaluating AIGIs. Specifically, A-Bench is organized under two key principles: 1) Emphasizing both high-level semantic understanding and low-level visual quality perception to address the intricate demands of AIGIs. 2) Various generative models are utilized for AIGI creation, and various LMMs are employed for evaluation, which ensures a comprehensive validation scope. Ultimately, 2,864 AIGIs from 16 text-to-image models are sampled, each paired with question-answers annotated by human experts. We hope that A-Bench will significantly enhance the evaluation process and promote the generation quality for AIGIs.

581. Null Counterfactual Factor Interactions for Goal-Conditioned Reinforcement Learning

链接：https://iclr.cc/virtual/2025/poster/31106 abstract： Hindsight relabeling is a powerful tool for overcoming sparsity in goal-conditioned reinforcement learning (GCRL), especially in certain domains such as navigation and locomotion. However, hindsight relabeling can struggle in object-centric domains. For example, suppose that the goal space consists of a robotic arm pushing a particular target block to a goal location. In this case, hindsight relabeling will give high rewards to any trajectory that does not interact with the block. However, these behaviors are only useful when the object is already at the goal---an extremely rare case in practice. A dataset dominated by these kinds of trajectories can complicate learning and lead to failures. In object-centric domains, one key intuition is that meaningful trajectories are often characterized by object-object interactions such as pushing the block with the gripper. To leverage this intuition, we introduce Hindsight Relabeling using Interactions (HInt), which combines interactions with hindsight relabeling to improve the sample efficiency of downstream RL. However, interactions do not have a consensus statistical definition that is tractable for downstream GCRL. Therefore, we propose a definition of interactions based on the concept of *null counterfactual*: a cause object is interacting with a target object if, in a world where the cause object did not exist, the target object would have different transition dynamics. We leverage this definition to infer interactions in Null Counterfactual Interaction Inference (NCII), which uses a ``nulling" operation with a learned model to simulate absences and infer interactions. We demonstrate that NCII is able to achieve significantly improved interaction inference accuracy in both simple linear dynamics domains and dynamic robotic domains in Robosuite, Robot Air Hockey, and Franka Kitchen. Furthermore, we demonstrate that HInt improves sample efficiency by up to $4\times$ in these domains as goal-conditioned tasks.

582. Generating Physical Dynamics under Priors

链接：https://iclr.cc/virtual/2025/poster/28929 abstract： Generating physically feasible dynamics in a data-driven context is challenging, especially when adhering to physical priors expressed in specific equations or formulas. Existing methodologies often overlook the integration of "physical priors", resulting in violation of basic physical laws and suboptimal performance. In this paper, we introduce a novel framework that seamlessly incorporates physical priors into diffusion-based generative models to address this limitation. Our approach leverages two categories of priors: 1) distributional priors, such as roto-translational invariance, and 2) physical feasibility priors, including energy and momentum conservation laws and PDE constraints. By embedding these priors into the generative process, our method can efficiently generate physically realistic dynamics, encompassing trajectories and flows. Empirical evaluations demonstrate that our method produces high-quality dynamics across a diverse array of physical phenomena with remarkable robustness, underscoring its potential to advance data-driven studies in AI4Physics. Our contributions signify a substantial advancement in the field of generative modeling, offering a robust solution to generate accurate and physically consistent dynamics.

583. Matryoshka Multimodal Models

链接：https://iclr.cc/virtual/2025/poster/29460 abstract： Large Multimodal Models (LMMs) such as LLaVA have shown strong performance in visual-linguistic reasoning. These models first embed images into a fixed large number of visual tokens and then feed them into a Large Language Model (LLM). However, this design causes an excessive number of tokens for dense visual scenarios such as high-resolution images and videos, leading to great inefficiency. While token pruning/merging methods do exist, they produce a single length output for each image and do not afford flexibility in trading off information density v.s. efficiency. Inspired by the concept of Matryoshka Dolls, we propose : Matryoshka Multimodal Models, which learns to represent visual content as nested sets of visual tokens that capture information across multiple coarse-to-fine granularities. Our approach offers several unique benefits for LMMs: (1) One can explicitly control the visual granularity per test instance during inference, e.g. , adjusting the number of tokens used to represent an image based on the anticipated complexity or simplicity of the content; (2) provides a framework for analyzing the granularity needed for existing datasets, where we find that COCO-style benchmarks only need around 9 visual tokens to obtain accuracy similar to that of using all 576 tokens; (3) Our approach provides a foundation to explore the best trade-off between performance and visual token length at sample level, where our investigation reveals that a large gap exists between the oracle upper bound and current fixed-scale representations.

584. TDDBench: A Benchmark for Training data detection

链接：https://iclr.cc/virtual/2025/poster/28739 abstract： Training Data Detection (TDD) is a task aimed at determining whether a specific data instance is used to train a machine learning model. In the computer security literature, TDD is also referred to as Membership Inference Attack (MIA). Given its potential to assess the risks of training data breaches, ensure copyright authentication, and verify model unlearning, TDD has garnered significant attention in recent years, leading to the development of numerous methods. Despite these advancements, there is no comprehensive benchmark to thoroughly evaluate the effectiveness of TDD methods.In this work, we introduce TDDBench, which consists of 13 datasets spanning three data modalities: image, tabular, and text. We benchmark 21 different TDD methods across four detection paradigms and evaluate their performance from five perspectives: average detection performance, best detection performance, memory consumption, and computational efficiency in both time and memory. With TDDBench, researchers can identify bottlenecks and areas for improvement in TDD algorithms, while practitioners can make informed trade-offs between effectiveness and efficiency when selecting TDD algorithms for specific use cases. Our extensive experiments also reveal the generally unsatisfactory performance of TDD algorithms across different datasets. To enhance accessibility and reproducibility, we open-source TDDBench for the research community at https://github.com/zzh9568/TDDBench.

585. GPS: A Probabilistic Distributional Similarity with Gumbel Priors for Set-to-Set Matching

链接：https://iclr.cc/virtual/2025/poster/29496 abstract： Set-to-set matching aims to identify correspondences between two sets of unordered items by minimizing a distance metric or maximizing a similarity measure. Traditional metrics, such as Chamfer Distance (CD) and Earth Mover's Distance (EMD), are widely used for this purpose but often suffer from limitations like suboptimal performance in terms of accuracy and robustness, or high computational costs - or both. In this paper, we propose a novel, simple yet effective set-to-set matching similarity measure, GPS, based on Gumbel prior distributions. These distributions are typically used to model the extrema of samples drawn from various distributions. Our approach is motivated by the observation that the distributions of minimum distances from CD, as encountered in real world applications such as point cloud completion, can be accurately modeled using Gumbel distributions. We validate our method on tasks like few-shot image classification and 3D point cloud completion, demonstrating significant improvements over state of-the-art loss functions across several benchmark datasets. Our demo code is publicly available at https://github.com/Zhang-VISLab/ICLR2025-GPS

586. RepoGraph: Enhancing AI Software Engineering with Repository-level Code Graph

链接：https://iclr.cc/virtual/2025/poster/28957 abstract： Large Language Models (LLMs) excel in code generation yet struggle with modern AI software engineering tasks. Unlike traditional function-level or file-level coding tasks, AI software engineering requires not only basic coding proficiency but also advanced skills in managing and interacting with code repositories. However, existing methods often overlook the need for repository-level code understanding, which is crucial for accurately grasping the broader context and developing effective solutions. On this basis, we present RepoGraph, a plug-in module that manages a repository-level structure for modern AI software engineering solutions. RepoGraph offers the desired guidance and serves as a repository-wide navigation for AI software engineers. We evaluate RepoGraph on the SWE-bench by plugging it into four different methods of two lines of approaches, where RepoGraph substantially boosts the performance of all systems, leading to a new state-of-the-art among open-source frameworks. Our analyses also demonstrate the extensibility and flexibility of RepoGraph by testing on another repo-level coding benchmark, CrossCodeEval. Our code is available at https://github.com/ozyyshr/RepoGraph.

587. BigCodeBench: Benchmarking Code Generation with Diverse Function Calls and Complex Instructions

链接：https://iclr.cc/virtual/2025/poster/29245 abstract： Task automation has been greatly empowered by the recent advances in Large Language Models (LLMs) via Python code, where the tasks range from software engineering development to general-purpose reasoning. While current benchmarks have shown that LLMs can solve tasks using programs like human developers, the majority of their evaluations are limited to short and self-contained algorithmic tasks or standalone function calls. Solving challenging and practical tasks requires the capability of utilizing diverse function calls as tools to efficiently implement functionalities like data analysis and web development. In addition, using multiple tools to solve a task needs compositional reasoning by accurately understanding complex instructions. Fulfilling both of these characteristics can pose a great challenge for LLMs. To assess how well LLMs can solve challenging and practical tasks via programs, we introduce BigCodeBench, a benchmark that challenges LLMs to invoke multiple function calls as tools from 139 libraries and 7 domains for 1,140 fine-grained tasks. To evaluate LLMs rigorously, each task encompasses 5.6 test cases with an average branch coverage of 99%. In addition, we propose a natural-language-oriented variant of BigCodeBench, BigCodeBench-Instruct, that automatically transforms the original docstrings into short instructions containing only essential information. Our extensive evaluation of 60 LLMs shows that LLMs are not yet capable of following complex instructions to use function calls precisely, with scores up to 60%, significantly lower than the human performance of 97%. The results underscore the need for further advancements in this area.

588. I^2AM: Interpreting Image-to-Image Latent Diffusion Models via Bi-Attribution Maps

链接：https://iclr.cc/virtual/2025/poster/29121 abstract： Large-scale diffusion models have made significant advances in image generation, particularly through cross-attention mechanisms. While cross-attention has been well-studied in text-to-image tasks, their interpretability in image-to-image (I2I) diffusion models remains underexplored. This paper introduces Image-to-Image Attribution Maps $(\textbf{I}^2\textbf{AM})$, a method that enhances the interpretability of I2I models by visualizing bidirectional attribution maps, from the reference image to the generated image and vice versa. I^2AM aggregates cross-attention scores across time steps, attention heads, and layers, offering insights into how critical features are transferred between images. We demonstrate the effectiveness of I^2AM across object detection, inpainting, and super-resolution tasks. Our results demonstrate that I^2AM successfully identifies key regions responsible for generating the output, even in complex scenes. Additionally, we introduce the Inpainting Mask Attention Consistency Score (IMACS) as a novel evaluation metric to assess the alignment between attribution maps and inpainting masks, which correlates strongly with existing performance metrics. Through extensive experiments, we show that I^2AM enables model debugging and refinement, providing practical tools for improving I2I model's performance and interpretability.

589. Monet: Mixture of Monosemantic Experts for Transformers

链接：https://iclr.cc/virtual/2025/poster/31204 abstract： Understanding the internal computations of large language models (LLMs) is crucial for aligning them with human values and preventing undesirable behaviors like toxic content generation. However, mechanistic interpretability is hindered by polysemanticity—where individual neurons respond to multiple, unrelated concepts. While Sparse Autoencoders (SAEs) have attempted to disentangle these features through sparse dictionary learning, they have compromised LLM performance due to reliance on post-hoc reconstruction loss. To address this issue, we introduce Mixture of Monosemantic Experts for Transformers (Monet) architecture, which incorporates sparse dictionary learning directly into end-to-end Mixture-of-Experts pretraining. Our novel expert decomposition method enables scaling the expert count to 262,144 per layer while total parameters scale proportionally to the square root of the number of experts. Our analyses demonstrate mutual exclusivity of knowledge across experts and showcase the parametric knowledge encapsulated within individual experts. Moreover, Monet allows knowledge manipulation over domains, languages, and toxicity mitigation without degrading general performance. Our pursuit of transparent LLMs highlights the potential of scaling expert counts to enhance mechanistic interpretability and directly resect the internal knowledge to fundamentally adjust model behavior.

590. Refine-by-Align: Reference-Guided Artifacts Refinement through Semantic Alignment

链接：https://iclr.cc/virtual/2025/poster/30475 abstract： Personalized image generation has emerged from the recent advancements in generative models. However, these generated personalized images often suffer from localized artifacts such as incorrect logos, reducing fidelity and fine-grained identity details of the generated results. Furthermore, there is little prior work tackling this problem. To help improve these identity details in the personalized image generation, we introduce a new task: reference-guided artifacts refinement. We present Refine-by-Align, a first-of-its-kind model that employs a diffusion-based framework to address this challenge. Our model consists of two stages: Alignment Stage and Refinement Stage, which share weights of a unified neural network model. Given a generated image, a masked artifact region, and a reference image, the alignment stage identifies and extracts the corresponding regional features in the reference, which are then used by the refinement stage to fix the artifacts. Our model-agnostic pipeline requires no test-time tuning or optimization. It automatically enhances image fidelity and reference identity in the generated image, generalizing well to existing models on various tasks including but not limited to customization, generative compositing, view synthesis, and virtual try-on. Extensive experiments and comparisons demonstrate that our pipeline greatly pushes the boundary of fine details in the image synthesis models.

591. Aligned Datasets Improve Detection of Latent Diffusion-Generated Images

链接：https://iclr.cc/virtual/2025/poster/28964 abstract： As latent diffusion models (LDMs) democratize image generation capabilities, there is a growing need to detect fake images. A good detector should focus on the generative model’s fingerprints while ignoring image properties such as semantic content, resolution, file format, etc. Fake image detectors are usually built in a data-driven way, where a model is trained to separate real from fake images. Existing works primarily investigate network architecture choices and training recipes. In this work, we argue that in addition to these algorithmic choices, we also require a well-aligned dataset of real/fake images to train a robust detector. For the family of LDMs, we propose a very simple way to achieve this: we reconstruct all the real images using the LDM's autoencoder, without any denoising operation. We then train a model to separate these real images from their reconstructions. The fakes created this way are extremely similar to the real ones in almost every aspect (e.g., size, aspect ratio, semantic content), which forces the model to look for the LDM decoder's artifacts. We empirically show that this way of creating aligned real/fake datasets, which also sidesteps the computationally expensive denoising process, helps in building a detector that focuses less on spurious correlations, something that a very popular existing method is susceptible to. Finally, to demonstrate the effectivenss of dataset alignment, we build a detector using images that are not natural objects, and present promising results. Overall, our work identifies the subtle but significant issues that arise when training a fake image detector and proposes a simple and inexpensive solution to address these problems.

592. Group-robust Sample Reweighting for Subpopulation Shifts via Influence Functions

链接：https://iclr.cc/virtual/2025/poster/29169 abstract： Machine learning models often have uneven performance among subpopulations(a.k.a., groups) in the data distributions. This poses a significant challenge for themodels to generalize when the proportions of the groups shift during deployment.To improve robustness to such shifts, existing approaches have developed strategiesthat train models or perform hyperparameter tuning using the group-labeled datato minimize the worst-case loss over groups. However, a non-trivial amount ofhigh-quality labels is often required to obtain noticeable improvements. Giventhe costliness of the labels, we propose to adopt a different paradigm to enhancegroup label efficiency: utilizing the group-labeled data as a target set to optimizethe weights of other group-unlabeled data. We introduce Group-robust SampleReweighting (GSR), a two-stage approach that first learns the representations fromgroup-unlabeled data, and then tinkers the model by iteratively retraining its lastlayer on the reweighted data using influence functions. Our GSR is theoreticallysound, practically lightweight, and effective in improving the robustness to sub-population shifts. In particular, GSR outperforms the previous state-of-the-artapproaches that require the same amount or even more group labels. Our code isavailable at https://github.com/qiaoruiyt/GSR.

593. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans?

链接：https://iclr.cc/virtual/2025/poster/28598 abstract： Comprehensive evaluation of Multimodal Large Language Models (MLLMs) has recently garnered widespread attention in the research community. However, we observe that existing benchmarks present several common barriers that make it difficult to measure the significant challenges that models face in the real world, including: 1) small data scale leads to a large performance variance; 2) reliance on model-based annotations results in restricted data quality; 3) insufficient task difficulty, especially caused by the limited image resolution. To tackle these issues, we introduce MME-RealWorld. Specifically, we collect more than 300 K images from public datasets and the Internet, filtering $13,366$ high-quality images for annotation. This involves the efforts of professional 25 annotators and 7 experts in MLLMs, contributing to $29,429$ question-answer pairs that cover 43 subtasks across 5 real-world scenarios, extremely challenging even for humans. As far as we know, **MME-RealWorld is the largest manually annotated benchmark to date, featuring the highest resolution and a targeted focus on real-world applications**. We further conduct a thorough evaluation involving 29 prominent MLLMs, such as GPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Our results show that even the most advanced models struggle with our benchmarks, where none of them reach 60\% accuracy. The challenges of perceiving high-resolution images and understanding complex real-world scenarios remain urgent issues to be addressed. The data and evaluation code are released in our Project Page.

594. xFinder: Large Language Models as Automated Evaluators for Reliable Evaluation

链接：https://iclr.cc/virtual/2025/poster/30821 abstract： The continuous advancement of large language models (LLMs) has brought increasing attention to the critical issue of developing fair and reliable methods for evaluating their performance. Particularly, the emergence of cheating phenomena, such as test set leakage and prompt format overfitting, poses significant challenges to the reliable evaluation of LLMs. As evaluation frameworks commonly use Regular Expression (RegEx) for answer extraction, models may adjust their responses to fit formats easily handled by RegEx. Nevertheless, the key answer extraction module based on RegEx frequently suffers from extraction errors. Furthermore, recent studies proposing fine-tuned LLMs as judge models for automated evaluation face challenges in terms of generalization ability and fairness. This paper comprehensively analyzes the entire LLM evaluation chain and demonstrates that optimizing the key answer extraction module improves extraction accuracy and enhances evaluation reliability. Our findings suggest that improving the key answer extraction module can lead to higher judgment accuracy and improved evaluation efficiency compared to the judge models. To address these issues, we propose xFinder, a novel evaluator for answer extraction and matching in LLM evaluation. As part of this process, we create a specialized dataset, the Key Answer Finder (KAF) dataset, to ensure effective model training and evaluation. Generalization tests and real-world evaluations show that the smallest xFinder model, with only 500 million parameters, achieves an average extraction accuracy of 93.42\%. In contrast, RegEx accuracy in the best evaluation framework is 74.38\%. The final judgment accuracy of xFinder reaches 97.61\%, outperforming existing evaluation frameworks and judge models.

595. CirT: Global Subseasonal-to-Seasonal Forecasting with Geometry-inspired Transformer

链接：https://iclr.cc/virtual/2025/poster/29244 abstract： Accurate Subseasonal-to-Seasonal (S2S) climate forecasting is pivotal for decision-making including agriculture planning and disaster preparedness but is known to be challenging due to its chaotic nature. Although recent data-driven models have shown promising results, their performance is limited by inadequate consideration of geometric inductive biases. Usually, they treat the spherical weather data as planar images, resulting in an inaccurate representation of locations and spatial relations. In this work, we propose the geometric-inspired Circular Transformer (CirT) to model the cyclic characteristic of the graticule, consisting of two key designs: (1) Decomposing the weather data by latitude into circular patches that serve as input tokens to the Transformer; (2) Leveraging Fourier transform in

self-attention to capture the global information and model the spatial periodicity. Extensive experiments on the Earth Reanalysis 5 (ERA5) reanalysis dataset demonstrate our model yields a significant improvement over the advanced data-driven models, including PanguWeather and GraphCast, as well as skillful ECMWF systems. Additionally, we empirically show the effectiveness of our model designs and high-quality prediction over spatial and temporal dimensions.

596. SimXRD-4M: Big Simulated X-ray Diffraction Data and Crystal Symmetry Classification Benchmark

链接：https://iclr.cc/virtual/2025/poster/28452 abstract： Powder X-ray diffraction (XRD) patterns are highly effective for crystal identification and play a pivotal role in materials discovery. While machine learning (ML) has advanced the analysis of powder XRD patterns, progress has been constrained by the limited availability of training data and established benchmarks. To address this, we introduce SimXRD, the largest open-source simulated XRD pattern dataset to date, aimed at accelerating the development of crystallographic informatics. We developed a novel XRD simulation method that incorporates comprehensive physical interactions, resulting in a high-fidelity database. SimXRD comprises 4,065,346 simulated powder XRD patterns, representing 119,569 unique crystal structures under 33 simulated conditions that reflect real-world variations. We benchmark 21 sequence models in both in-library and out-of-library scenarios and analyze the impact of class imbalance in long-tailed crystal label distributions. Remarkably, we find that: (1) current neural networks struggle with classifying low-frequency crystals, particularly in out-of-library situations; (2) models trained on SimXRD can generalize to real experimental data.

597. MuPT: A Generative Symbolic Music Pretrained Transformer

链接：https://iclr.cc/virtual/2025/poster/28712 abstract： In this paper, we explore the application of Large Language Models (LLMs) to the pre-training of music. While the prevalent use of MIDI in music modeling is well-established, our findings suggest that LLMs are inherently more compatible with ABC Notation, which aligns more closely with their design and strengths, thereby enhancing the model's performance in musical composition.To address the challenges associated with misaligned measures from different tracks during generation, we propose the development of a $\underline{S}$ynchronized $\underline{M}$ulti-$\underline{T}$rack ABC Notation ($\textbf{SMT-ABC Notation}$), which aims to preserve coherence across multiple musical tracks. Our contributions include a series of models capable of handling up to 8192 tokens, covering 90\% of the symbolic music data in our training set. Furthermore, we explore the implications of the $\underline{S}$ymbolic $\underline{M}$usic $\underline{S}$caling Law ($\textbf{SMS Law}$) on model performance. The results indicate a promising research direction in music generation, offering extensive resources for further research through our open-source contributions.

598. Shapley-Guided Utility Learning for Effective Graph Inference Data Valuation

链接：https://iclr.cc/virtual/2025/poster/30761 abstract： Graph Neural Networks (GNNs) have demonstrated remarkable performance in various graph-based machine learning tasks, yet evaluating the importance of neighbors of testing nodes remains largely unexplored due to the challenge of assessing data importance without test labels. To address this gap, we propose Shapley-Guided Utility Learning (SGUL), a novel framework for graph inference data valuation. SGUL innovatively combines transferable data-specific and modelspecific features to approximate test accuracy without relying on ground truth labels. By incorporating Shapley values as a preprocessing step and using feature Shapley values as input, our method enables direct optimization of Shapley value prediction while reducing computational demands. SGUL overcomes key limitations of existing methods, including poor generalization to unseen test-time structures and indirect optimization. Experiments on diverse graph datasets demonstrate that SGUL consistently outperforms existing baselines in both inductive and transductive settings. SGUL offers an effective, efficient, and interpretable approach for quantifying the value of test-time neighbors.

599. A Solvable Attention for Neural Scaling Laws

链接：https://iclr.cc/virtual/2025/poster/27834 abstract： Transformers and many other deep learning models are empirically shown to predictably enhance their performance as a power law in training time, model size, or the number of training data points, which is termed as the neural scaling law. This paper studies this intriguing phenomenon particularly for the transformer architecture in theoretical setups. Specifically, we propose a framework for linear self-attention, the underpinning block of transformer without softmax, to learn in an in-context manner, where the corresponding learning dynamics is modeled as a non-linear ordinary differential equation (ODE) system. Furthermore, we establish a procedure to derive a tractable approximate solution for this ODE system by reformulating it as a Riccati equation, which allows us to precisely characterize neural scaling laws for linear self-attention with training time, model size, data size, and the optimal compute. In addition, we reveal that the linear self-attention shares similar neural scaling laws with several other architectures when the context sequence length of the in-context learning is fixed, otherwise it would exhibit a different scaling law of training time.

600. DyCAST: Learning Dynamic Causal Structure from Time Series

链接：https://iclr.cc/virtual/2025/poster/29348 abstract： Understanding the dynamics of causal structures is crucial for uncovering the underlying processes in time series data. Previous approaches rely on static assumptions, where contemporaneous and time-lagged dependencies are assumed to have invariant topological structures. However, these models

fail to capture the evolving causal relationship between variables when the underlying process exhibits such dynamics. To address this limitation, we propose DyCAST, a novel framework designed to learn dynamic causal structures in time series using Neural Ordinary Differential Equations (Neural ODEs). The key innovation lies in modeling the temporal dynamics of the contemporaneous structure, drawing inspiration from recent advances in Neural ODEs on constrained manifolds. We reformulate the task of learning causal structures at each time step as solving the solution trajectory of a Neural ODE on the directed acyclic graph (DAG) manifold. To accommodate high-dimensional causal structures, we extend DyCAST by learning the temporal dynamics of the hidden state for contemporaneous causal structure. Experiments on both synthetic and real-world datasets demonstrate that DyCAST achieves superior or comparable performance compared to existing causal discovery models.